Stability of Executive Functioning Measures in 8–17-Year-Old Children With Unilateral Cerebral Palsy

Adina M. Piovesana, Stephanie Ross, Koa Whittingham, Robert S. Ware & Roslyn N. Boyd

Stability of Executive Functioning Measures in 8–17-Year-Old Children With Unilateral Cerebral Palsy

Adina M. Piovesana[1], Stephanie Ross[1], Koa Whittingham[1,2], Robert S. Ware[3,4], and Roslyn N. Boyd[1]

[1]Queensland Cerebral Palsy and Rehabilitation Research Centre, School of Medicine, The University of Queensland, Brisbane, QLD 4029, Australia
[2]School of Psychology, The University of Queensland, Brisbane, QLD 4029, Australia
[3]School of Population Health, The University of Queensland, Brisbane, QLD 4029, Australia
[4]Queensland Children's Medical Research Institute, The University of Queensland, Brisbane, QLD 4029, Australia

The study investigated the stability of executive functioning (EF) measures in children and adolescents aged 8–17 years with unilateral cerebral palsy (CP). Here 44 participants with unilateral CP (mean age = 11 years, 11 months; Manual Abilities Classification Scale Level I = 6 and Level II = 37; Gross Motor Function Classification Scale Level I = 22 and Level II = 22) were randomized into the wait-list control group of a large randomized controlled trial. Participants had baseline testing with Wechsler Intelligence Scale for Children – Fourth Edition Short Form (WISC-IV-SF) and Delis-Kaplan Executive Function System (D-KEFS) subtests. Parents completed the Behavior Rating Inventory of Executive Functioning (BRIEF). Participants were reassessed 20 ±2 weeks later with a shortened test battery including the D-KEFS subtests; Digit Span, Coding and Symbol Search (WISC-IV); and BRIEF. Pearson's test–retest reliabilities and Reliable change scores were calculated. Results indicated excellent to fair test–retest reliabilities ($r = 0.91$–0.74) for all measures except Digit Span Backwards ($r = 0.62$), Inhibition ($r = 0.69$), and Initiate ($r = 0.68$). Reliable change scores applying 90% confidence intervals for estimating reliable change while accounting for practice effects were provided for all measures. The data support the stability of EF measures in this population.

INTRODUCTION

Cerebral palsy (CP) is the major cause of childhood disability worldwide, with an estimated incidence of 2.9 per 1000 live births (Himmelmann & Uvebrant, 2014). The incidence rates are comparable in Australia, with 2.2 per 1000 live births (Stanley, Blair, & Alberman, 2000; The Australian Cerebral Palsy Register Group, 2013). Unilateral CP or congenital hemiplegia, where one half of the body has impairment with voluntary movement, accounts for approximately 44% of children born with CP worldwide and 39% in Australia (Himmelmann & Uvebrant, 2014; The Australian Cerebral Palsy Register Group, 2013). Cerebral palsy is an umbrella term describing a group of permanent disorders of movement, posture, and/or motor function that occur

as a result of non-progressive abnormalities to the infant or fetal developing brain (Rosenbaum, Paneth, Leviton, Goldstein, & Bax, 2006). Although CP is primarily a movement disorder, children and adolescents may have other associated diagnoses such as epilepsy, and hearing, vision, and speech/language impairments (Bax et al., 2005; Guzzetta, Mercuri, & Cioni, 2001; Kennes et al., 2007; Krigger, 2007; Pirila et al., 2007; The Australian Cerebral Palsy Register Group, 2013). Furthermore, the impact of brain insult on intellectual impairments and other cognitive difficulties is gaining recognition (Odding, Roebroeck, & Stan, 2006). Recent research has found that children with an early brain insult (EBI) are at increased risk of having an executive dysfunction compared to their typically developing peers (Anderson, Spencer-Smith, & Wood, 2011).

Executive functioning (EF) is an umbrella term used to describe higher-order cognitive abilities and control processes such as inhibition of inappropriate or automatic responses, initiation and planning of behavior, multi-tasking, cognitive flexibility, judgment and decision making, and monitoring performance (Collette, Hogge, Salmon, & Van, 2006). These functions "are responsible for purposeful, goal-directed activity, and include the highest level of human functioning such as intellect, thought, self-control, and social interaction" (Lezak, 1995, p. 42). In 2002 Anderson proposed a conceptual framework for EF in children and adolescents. Other models have conceptualized EF in children and adolescents as a single construct (Della Sala, Gray, Spinnler, & Trivello, 1998), multiple inter-related executive processes (Alexander & Stuss, 2000; Stuss & Alexander, 2000), or a supervisory system (Duncan, 1995). Unlike these models, the Anderson model considers EF as an overall control system which is comprised of four distinct yet integrated components: (1) attentional control, (2) cognitive flexibility, (3) goal-setting, and (4) information processing. The Anderson model is an empirically validated model, developed using factor-analytic studies and was specifically designed for a developmental context. It has also been used in other developmental neuropsychological studies that are concerned with EF and children with CP and as such is deemed an appropriate model for the study (Bodimeade, Whittingham, Lloyd, & Boyd, 2013; Whittingham, Bodimeade, Lloyd, & Boyd, 2014), and those with early brain insult (Anderson et al., 2010).

Children with unilateral CP have consistently been found to have significantly decreased EF in daily life compared to typically developing children (TDC; Bottcher, Flachs, & Uldall, 2010; Pirila, van der Meere, Rantanen, Jokiluoma, & Eriksson, 2011; Straub & Obrzut, 2009; Whittingham et al., 2014). In one study 46 children with unilateral CP were compared with 20 TDC on a range of neuropsychological measures specifically selected to assess the four EF components as outlined by Anderson (2002; see also Bodimeade et al., 2013). Results indicated that children with unilateral CP, regardless of laterality, performed significantly worse than TDC on measures of attentional control, cognitive flexibility, goal setting, and information processing, with the exception of inhibition/switching within attentional control where right-sided unilateral CP children (i.e., left-sided lesion) had more errors (Bodimeade et al., 2013). In a further analysis using the same cohort, children with unilateral CP were found to display significantly worse behavioral manifestations of EF in daily life as measured by the Behavioral Rating Inventory of Executive Functioning (BRIEF; Gioia, Isquith, Guy, & Kenworthy, 2000) as well as increased hyperactivity and conduct problems as measured by the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1999) when compared to a group of TDC (Whittingham et al., 2014). The increased risk of EF difficulties in

everyday life, conduct problems, and hyperactivity was found to be partly explained by decreased cognitive EF abilities as measured by neuropsychological assessment. Overall, children and adolescents with CP have been found to have significant impairments in visual and auditory attention, working memory, planning, inhibition, and response time (Anderson et al., 2010; Burnett, Scratch, & Anderson, 2013; Edgin et al., 2008; Pirila et al., 2011).

It is therefore crucial to examine test–retest reliability of EF assessments in the CP population as a way to evaluate stability of functioning (i.e., whether any observed scores are both significantly and clinically meaningful) and the efficacy of rehabilitation programs that target cognitive functioning. No neuropsychological test to date is completely reproducible; that is, all have some degree of error associated with administration. Error can be introduced from the test itself (i.e., measurement error), from repeated administrations (i.e., practice effects), and also from the patient and environmental factors (Barr, 2002). Patient factors may include the client's behavior, psychological status, concentration, and motivation levels, whereas the testing environment can include external noise and other distractors. In the neuropsychology literature, a reliable change index (RCI) "is a statistical method for developing empirically derived cut-offs that can be used for evaluating meaningful differences in test scores, independent of psychometric issues, such as practice effects and other sources of variance" (Barr, 2002, p. 302). In other words, RCIs determine the amount of change needed to be clinically meaningful, after accounting for the sources of error associated with repeated neuropsychological administrations (Hinton-Bayre, 2010).

Reliable Change Indices can be calculated using a variety of formulae, however there is considerable and ongoing debate regarding the efficacy and application of these. For a comprehensive review see Hinton-Bayre (2010). On the most basic level, all RCIs are modeled on the following formula where X_2 is the retest score X_1 is the initial test score and SE is the standard error of measurement of the difference between the two test scores (Jacobson & Truax, 1991):

$$RC_1 = X_2 - X_1 / SE$$

This basic structure, however, does not account for the error associated with practice effects. If practice effects are found in the measures being used, then clinicians can use an adjusted RCI developed by Iverson and Green (2001) as it accounts for this error by applying a uniform correction for practice based on mean practice effect seen in a control group. This adapted formula where X_2 is the retest score, X_1 is the initial test score, and S_{diff} is the standard error of difference equals:

$$RC_2 = X_2 - X_1 / S_{diff}$$

The S_{diff} estimates the measurement errors associated with the initial and retest scores and is calculated by the following formula:

$$S_{diff} = \sqrt{SEM^2_1 + SEM^2_2}$$

(Square root of the sum of squared standard error of measurements (SEM) for initial and retest scores.)

Essentially, the S_{diff} "is used to create a confidence interval for the baseline–retest difference score" (Iverson, Lovell, & Collins, 2003, p. 462). The SEM is the amount of measurement error associated with a score and is obtained by the following formula:

$$\text{SEM} = \text{SD}\sqrt{1 - r_{12}}$$

where SD is the standard deviation of the sample and r_{xy} is the test–retest coefficient

The S_{diff} can be used to calculate 90% confidence intervals using the z-score transformation which helps a clinician determine whether a test performance represents real change above and beyond practice effects (Iverson & Green, 2001).

Results of stability studies should only be generalized to samples with similar characteristics, so it is necessary to obtain test–retest data for the CP population itself. Given the importance of testing EF in children with unilateral CP, there is a need to develop specific test–retest data and RCIs and/or reliability parameters specific to this population. To date there are no published studies that report test–retest stability or reliability for this population on EF tests. The aim of this current study is to investigate the test–retest stability of a range of EF measures in children aged 8–17 years with mild to moderate unilateral CP.

METHOD

Participants

Participants of this study were from the wait-list control group of a large Randomized Controlled Trial (RCT) looking at the effects of an online rehabilitation program (Boyd et al., 2013). Participants had been matched and randomized into the wait-list control condition and were to receive 20 weeks of standard care before receiving the Move-It-To-Improve-It (Mitii[TM]) intervention. The Mitii[TM] intervention is an interactive web-based multi-modal training intervention that combines upper limb and occupational performance, physical activity, and cognitive training. Participants had a confirmed diagnosis of unilateral CP and were recruited from Queensland and New South Wales using a population-based research database comprising 1300 registered children with CP. Participants were included in the study if they had the ability to follow instructions, had mild to moderate congenital hemiplegia (as diagnosed by their treating physician), had a Gross Motor Function Classification Scale (GMFCS; Palisano et al., 2000) of I or II, a Manual Abilities Classification Scale (MACS; Eliasson et al., 2006) of I, II, or III and were between 8 and 18 years of age. Participants were excluded if they were involved in another exercise program, had had upper or lower limb surgery in the 6 months prior to recruitment, had unstable epilepsy, or had respiratory, cardiovascular, or other medical condition.

The GMFCS is a system that classifies children based on their ability to carry out self-initiated movements (e.g., sitting and walking) and is based on five ability levels. At the lowest level (Level I) the child is able to walk without limitations and is able to function in community settings with minimal functional difficulties. On the other hand, a child who has great difficulty with independent and voluntary movements, uses a wheelchair, and is typically fully dependent on carers would be classified as Level V. The GMFCS is highly correlated ($r = 0.91$) with the Gross Motor Function Measure

(GMFM; Palisano et al., 2000) and has high test–retest reliability ($r = 0.79$; Wood & Rosenbaum, 2000), and inter-rater reliability between professionals (kappa = 0.74; Liu, Thawinchain, Palizano, & Valvano, 1998). Similarly, MACS is a five-levelled system that classifies children based on their ability to handle objects in daily activities. A child at Level I would easily be able to handle objects, whereas a child at Level V has a severely limited ability to perform simple actions and handle objects. The MACS has demonstrated excellent inter-rater ($r = 0.97$) and intra-rater reliability ($r = 0.96$), and is highly correlated with the GMFCS ($r = 0.79$; Eliasson et al., 2006).

Materials

For the initial testing session, participants were administered the Wechsler Intelligence Scale for Children – Fourth Edition Short Form (WISC-IV-SF) developed by Crawford, Anderson, Rankin, and MacDonald (2010). The WISC-IV-SF provided a measure of intellectual functioning across the four domains of Verbal Comprehension (VCI), Perceptual Reasoning (PRI), Working Memory (WMI), and Processing Speed (PSI). The VCI, PRI, and PSI index scores have moderate to excellent levels of internal consistency ($\alpha = 0.88–0.96$) in TDC and are comparable to the full version of the WISC-IV (Crawford et al., 2010). The WMI index internal consistency ($\alpha = 0.87$) is marginally lower than the full-length equivalent ($\alpha = 0.92$; Wechsler, 2003) in TDC.

Participants were also administered tests specifically designed to measure the EF domains as outlined by Anderson (2002). Measures were selected based on ease and length of administration, test–retest reliability, had been previously used in the target population (Anderson et al., 2010; Bodimeade et al., 2013) and because the measures have also been previously used in the domains of the Anderson EF model (2002). Measures included the Color-Word interference test, TMT (cognitive flexibility), and Tower Test (goal setting) subtests of the D-KEFS (Delis, Kaplan, & Kramer, 2001). Although these subtests were administered in their entirety, only specific conditions and/or scores specifically measured EF. The Inhibition condition from the Color-Word interference test was used as a measure of attentional control. This condition was also utilized by Anderson et al. (2010) in order to assess attentional control. Participants were required to name the ink color in which the names of colors were printed (e.g., say "green" for the word "red" that was printed in green ink) while being timed. This measure has an excellent test–retest reliability ($r = 0.90$) in the standardization sample of TDC (Delis et al., 2001).

The Switching condition from the TMT was used as a measure of cognitive flexibility as participants were required to connect numbers and letters in order (e.g., "1-A-2-B-3-C"). While this condition has a poor documented test–retest reliability ($r = 0.20$) in 8–19-year-old TDC (Delis et al., 2001), the D-KEFS TMT subtest is easy to administer, with clear instructions, and has been previously used to measure this component of EF (Anderson et al., 2010). Goal setting was assessed using the Tower Test. Participants were required to move five different-sized disks across three pegs in order to build a target tower in a specific time limit and using the fewest number of moves possible. This measure has an adequate test–retest reliability ($r = 0.51$) and moderate internal consistency ($\alpha = 0.63$) for 8–19-year-old TDC (Delis et al., 2001) and has been used in previous studies to measure this subdomain of EF (Anderson et al., 2010).

The Digit Span Backwards subtest of the WISC-IV was also used to assess cognitive flexibility (Wechsler, 2003) as it requires the skills of working memory, divided attention, and shifting behavior. In Digit Span Backwards, participants were required to verbally repeat a string of numbers in the reverse order (e.g., if "1-4-5", then correct answer is "5-4-1"). It has good internal consistency ($\alpha = 0.80$) and adequate test–retest reliability ($r = 0.74$) in TDC (Wechsler, 2003). The Coding and Symbol Search subtests of the WISC-IV were also used to assess processing speed (Wechsler, 2003). Coding and Symbol Search also have been shown to have adequate to good levels of internal consistency ($\alpha = 0.79$–0.82) and good test–retest reliabilities ($r = 0.80$–0.81; Wechsler, 2003). In Coding, participants were required to match up and draw abstract geometric shapes with numbers from a key within a 2-minute time period. In Symbol Search, participants were required to visually scan for target symbols in groups of five symbols and indicate whether the target symbol appeared.

Parents or guardians completed the BRIEF (Gioia et al., 2000) which assesses behavioral manifestations of EF in daily life. The BRIEF has been shown to have good internal consistency ($\alpha = 0.80$–0.98) and high test–retest reliabilities across indexes ($r = 0.86$–0.92) in TDC (Gioia et al., 2000). Parents or guardians also completed an additional un-standardized questionnaire relating to demographic characteristics of the participant and family (Boyd et al., 2013).

For the second testing session, 20 weeks later, participants were administered a shortened battery of tests which included the D-KEFS subtests (i.e., Color-Word Inference Test, TMT, and Tower Test) and Digit Span, Coding, and Symbol Search of the WISC-IV. The BRIEF was completed by parents or guardians, as was an un-standardized questionnaire specifically designed by the research team to assess duration, frequency, and content of standard of care and concurrent therapies during the retest period.

Standard scores were calculated for all tests using age-appropriate normative data provided by test manuals. Subtests of the WISC-IV and WISC-IV-SF were transformed into scaled scores (mean $= 10$, $SD = 3$). The Full Scale Intelligence Quotient (FSIQ) and index scores (i.e., VCI, PRI, WMI, and PSI) calculated from the WISC-IV-SF were also transformed into scaled scores but had a mean of 100 and standard deviation of 15. Raw scores were used for the D-KEFS subtests when calculating Pearson's test–retest reliabilities in order to make comparisons with the normative data contained within the manual (Delis et al., 2001). However, RCI were calculated using scaled scores (mean $= 10$, $SD = 3$). For the BRIEF, raw scores were converted into T scores which had a mean on 50 and standard deviation of 10.

Procedure

Test–retest interval in this study was 20 ±2 weeks. While this retest interval is not conventional in published test–retest data (i.e., some measures report test–retest intervals of 2–3 weeks), it was chosen to reflect a typical clinical retest interval in neuropsychological testing where patients are usually retested at longer intervals. During the retest interval participants continued to receive standard care and were not involved in any concomitant treatments during the retest interval.

Assessments were carried out in a clinical setting at the University of Queensland in Brisbane. Two fully registered and experienced psychologists administered the battery of EF measures to the participants using standardized administration guidelines. Most participants were tested early to late morning to ensure full attention and concentration. For the retest, participants were not necessary administered the tests by the same psychologist as the initial session, however standardized administration and scoring was used.

Ethics approval was obtained from the Medical Ethics Committee of The University of Queensland (2011000608) and with the Children's Health Service Human Research Ethics Committee at the Royal Children's Hospital Brisbane (HREC/11/QRCH/35). The study was registered with the Australian New Zealand Clinical Trials Registry (ACTRN12611001174976). Written consent was obtained from parents or guardians and all participants over the age of 12. Consent was obtained by study coordinators and personnel, upon entering the clinical trial.

Statistical analyses

The standard error of measurement (SEM) was calculated using standard formula (de Vet, Terwee, Knol, & Bouter, 2006). Pearson test–retest correlation coefficients were also calculated to assess reliability. According to Cicchetti (1994) reliability coefficients less than 0.70 are considered unacceptable, between 0.70 and 0.79 are considered fair, between 0.80 and 0.89 are considered good, and reliabilities above 0.90 are considered excellent.

Statistically significant practice effects were evaluated using paired-samples t tests. Significant mean practice effects occur if the mean practice t-score (i.e., $M_Y–M_x$) fall outside the cut-off score of $±1.645$ (upper bound of the 90% confidence interval), where M_Y is the initial test mean; and M_x is the retest mean. Given the small sample size, a modified t score was used to determine significant change based on the degrees of freedom (df = N–1; Hinton-Bayre, 2012). The standard error of difference (S_{diffs}) and the 90% confidence interval for estimating change were used to estimate the minimal change in score needed for noticeable change above and beyond measurement error plus practice effects. All analyses were calculated using the Statistical Package for Social Sciences (version 22.0).

RESULTS

Demographic characteristics for the 44 participants are reported in Table 1. Gender was near equal with 21 (47.7%) males and the mean (SD) age was 11 years, 11 months (2 years, 6 months). The mean FSIQ for the study participants was 79.3 (±20.2) with a range of 42 to 116. The FSIQ mean and standard deviation (SD) have been determined based on the population-level data, and as such children with CP are likely to be more heterogeneous than children in the general population. It is not surprising that a larger SD was observed for FSIQ in this sample. Additional demographic characteristics including medical diagnoses and socio-economic information are also included in this table. The use and frequency of standard of care by participants between testing periods as measured by the un-standardized questionnaire completed by parents during

Table 1. Demographics of participants ($n = 44$)

Participant characteristics	
Age (years) Mean ± SD	11.96 ±2.47
Gender n (%)	
Male	21 (47.7%)
Female	23 (52.3%)
MACS distribution n (%)	
Level I	6 (14%)
Level II	37 (86%)
GMFCS n (%)	
Level I	22 (50%)
Level II	22 (50%)
FSIQ Mean ± SD	79.33 ± 20.23 ($n = 43$)
VCI	87.47 ± 16.76 ($n = 43$)
PRI	80.33 ± 18.12 ($n = 43$)
WMI	85.12 ± 16.05 ($n = 43$)
PSI	82.72 ± 19.68 ($n = 43$)
SEIFA Socio-Economic Disadvantage Decile	6.46 ±2.98
Diagnosis of Intellectual Disability	7 (15.9%)
Diagnosis of Learning Difficulties	9 (20.5%)
Diagnosis of Autism	1 (2.3%)
Diagnosis of ADHD	5 (11.4%)
Diagnosis of Vision Impairment	6 (13.6%)
Diagnosis of Hearing Impairment	3 (6.8%)
Diagnosis of Stable Epilepsy	7 (15.9%)

SD = Standard Deviation; % = percentage; MACS = Manual Ability Classification system; GMFCS = Gross Motor Function Classification System; n = sample size; FSIQ = Full Scale Intelligence Quotient; VCI = Verbal Comprehension Index; PRI = Perceptual Reasoning Index; WMI = Working Memory Index; PSI = Processing Speed Index; SEIFA = Socio-Economic Indexes for Areas; ADHD = Attention Deficit Hyperactivity Disorder.

the second testing session is reported in Table 2. Physiotherapy was the most utilized of the therapies with 31 participants ($n = 31$) receiving physiotherapy between testing periods. However this was infrequent with 8% receiving therapy once per term and 22% receiving therapy once per year.

Test–retest stability information for the WISC-IV and D-KEFS subtests and the BRIEF is presented in Table 3. Also contained in this table are the test–retest reliabilities of age-related TDC published in the test manuals for the measures. As mentioned, the test–retest reliabilities reported in the manual for the D-KEFS TMT, Color Word Interference Test, and Tower Test are based on the raw scores (completion time or total moves) and not on scaled scores. Test–retest reliabilities for the WISC-IV subtests and the BRIEF are calculated using standardized scores. For the purposes of comparison, test–retest reliabilities for the D-KEFS were also calculated using raw scores. Statistically significant differences between test–retest reliabilities for the current sample and those reported in the test manuals are also reported in Table 3. Pearson's test–retest reliabilities ranged from excellent to fair ($r = 0.91$–0.74) with the exception of Digit Span Backwards ($r = 0.62$), the Inhibition condition of the D-KEFS Color-Word Interference Test ($r = 0.69$), and the BRIEF Initiate ($r = 0.68$). Both of these measures demonstrated poor stability and reliability across the testing sessions.

Table 2. Use and frequency of concurrent therapies
(*n* = 44)

Concurrent therapies	*n* (%)
Physiotherapist	(*n* = 31, 71%)
More than once per fortnight	1 (1%)
Once a term	8 (7.8%)
Once per year	22 (21.6%)
Occupational Therapist	(*n* = 21, 48%)
Once per week	1 (1%)
More than once per fortnight	1 (1%)
Once a term	2 (2%)
Once per year	17 (16.7%)
Psychologist	(*n* = 11, 25%)
Once per month	1 (1%)
Once per term	6 (5.9%)
Once per year	4 (3.9%)
Orthotists/Prosthetics	(*n* = 18, 41%)
Once per week	1 (1%)
Once per term	1 (1%)
Once per year	16 (15.7%)
Pediatrician	(*n* = 24, 55%)
More than once per fortnight	1 (1%)
Once per term	2 (2%)
Once per year	21 (20.6%)

n = sample size.

Test–retest reliability was excellent for Color Naming (r = 0.91) condition of the Color-Word Interference Test, and the BRIEF Global Executive Composite (GEC; r = 0.90). Test–retest reliabilities were good for the Meta-Cognition Index (r = 0.82), Behavioral Regulation Index (r = 0.82), Working Memory (r = 0.82), Plan/Organize (r = 0.87), Organization of Materials (r = 0.86), and the Monitor (r = 0.89) subtests of the BRIEF; the Digit Span Forwards (r = 0.80), Digit Span Total (r = 0.80), Coding (r = 0.85), and Symbol Search (r = 0.85) subtests of the WISC-IV; Letter Sequencing (r = 0.82) and Number-Letter Sequencing (r = 0.81) of the D-KEFS TMT; and the Word Reading (r = 0.89) condition of the D-KEFS Color-Word Interference Test. Test–retest reliabilities were fair for the Number Sequencing (r = 0.77) and Motor Speed (r = 0.75) conditions of the D-KEFS TMT; the D-KEFS Tower Test Achievement Score (r = 0.74); and the Inhibit (r = 0.74), Shift (r = 0.74), and Emotional Control (r = 0.77) subtests of the BRIEF. Statistically significant differences between the test–retest reliabilities in the current sample and those reported in the manuals were found for Symbol Search (z = 2.5, p = .01); Letter Sequencing (z = 2.31, p = .02); Number Letter Sequencing (z = 4.18, p = .00); and BRIEF Monitor (z = 2.62, p = .01). The test–retest reliabilities for these measures were better than those reported in the manuals. A statistically significant difference was also found for the Inhibition subtest (z = –3.08, p = .00) however the test–retest reliability published in the manual was better than in the current sample.

Statistically significant mean practice effects (i.e., difference between the means at time one and time two assessments) using paired-samples t tests were found on TMT Letter Sequencing t(43) = 2.80, p = .08; TMT Letter Number Sequencing t(43) = 2.11,

Table 3. Test–retest reproducibility in children with unilateral cerebral palsy ($n = 44$) for neuropsychological measures

Measures	Time 1 $M \pm SD$	Time 2 $M \pm SD$	Mean Diff	SEM	r_{xy}	r_{xy}[a]	r_{xy} Diff.[b]
WISC-IV ($n = 43$)							
Digit Span Forwards	6.86 ± 3.08	7.49 ± 3.50	0.63	1.49	0.80	0.72*	$z = 1.12; p = .26$
Digit Span Backwards	7.72 ± 2.98	7.51 ± 2.69	−0.21	1.76	0.62	0.67*	$z = -0.5; p = .62$
Digit Span Total	6.98 ± 3.23	7.09 ± 3.06	0.12	1.41	0.80	0.81*	$z = -0.17; p = .87$
Coding	6.95 ± 3.40	7.19 ± 3.45	0.23	1.32	0.85	0.81*	$z = 0.76; p = .45$
Symbol Search	6.95 ± 4.07	7.58 ± 3.98	0.63	1.59	0.85	0.68*	$z = 2.5; p = .01$
D-KEFS TMT ($n = 44$)							
Number-Sequencing	58.52 ± 29.51	52.20 ± 24.54	−6.32	13.36	0.77	0.77*	$z = 0.0; p = 1$
Letter-Sequencing	70.32 ± 39.30	58.52 ± 35.37	−11.80	16.11	0.82	0.57*	$z = 2.31; p = .02$
Number-Letter Seq.	151.34 ± 64.28	131.84 ± 60.86	−19.50	27.23	0.81	0.20*	$z = 4.18; p = .00$
Motor Speed	39.20 ± 23.44	35.75 ± 20.55	−3.46	11.11	0.75	0.82*	$z = -0.83; p = .41$
D-KEFS Color-Word Test ($n = 44$)							
Word Reading	35.98 ± 14.38	35.07 ± 12.80	−0.91	4.72	0.89	0.77*	$z = 1.82; p = .07$
Color Naming	45.48 ± 14.30	46.00 ± 15.64	0.52	4.65	0.91	0.79*	$z = 2.07; p = .04$
Inhibition	94.05 ± 36.33	82.66 ± 31.21	−11.39	19.12	0.69	0.91*	$z = -3.08; p = .00$
D-KEFS Tower Test							
Achievement	12.25 ± 4.08	15.30 ± 4.89	3.05	2.35	0.74	0.51*	$z = -1.76; p = .08$
BRIEF ($n = 42$)							
Inhibit	59.12 ± 15.08	59.52 ± 15.14	0.41	7.68	0.74	0.84	$z = -1.67; p = .10$
Shift	60.93 ± 12.48	60.69 ± 14.03	−0.24	6.89	0.74	0.78	$z = -0.58; p = .56$
Emotional Control	59.45 ± 14.12	58.24 ± 14.04	−1.21	6.77	0.77	0.79	$z = -0.31; p = .76$
Initiate	59.76 ± 13.87	60.76 ± 11.99	1	6.63	0.68	0.80	$z = -1.66; p = .10$
WM	62.60 ± 12.06	64.48 ± 11.61	1.88	5.08	0.82	0.85	$z = -0.61; p = .54$
Plan/Organize	62.64 ± 12.11	64.02 ± 12.29	1.38	4.4	0.87	0.85	$z = 0.47; p = .64$
Org. of Materials	55.14 ± 9.63	56.43 ± 9.99	1.29	3.68	0.86	0.79	$z = 1.37; p = .17$
Monitor	59.64 ± 13.43	60.95 ± 12.47	1.31	4.31	0.89	0.76	$z = 2.62; p = .01$

(continued)

Table 3. (Continued)

Measures	Time 1 $M \pm SD$	Time 2 $M \pm SD$	Mean Diff.	SEM	r_{xy}	$r_{xy}{}^a$	r_{xy} Diff.[b]
BRI	61.43 ± 14.60	61.07 ± 13.87	−0.36	6.14	0.82	0.84	$z = -0.4; p = .69$
MCI	62.24 ± 11.51	64.81 ± 12.58	2.57	5.13	0.82	0.88	$z = -1.35; p = .18$
GEC	62.62 ± 13.07	63.79 ± 13.11	1.17	4.15	0.9	0.86	$z = 1.1; p = .27$

n = sample size; M = mean; Diff. = difference; SD = standard deviation; SEM = standard error of measurement; r_{xy} = Pearson's test–retest reliability correlation; WISC-IV = Wechsler Intelligence Scale for Children – Fourth Edition; z = z-score; p = probability value using two-tailed test of significance; D-KEFS = Delis-Kaplan Executive Function System; TMT = Trail Making Test; Seq. = sequence; BRIEF = Behavior Rating Inventory of Executive Function; WM = working memory; Org. = Organization; BRI = Behavioral Rating Index; MCI = Metacognition Index; GEC = Global Executive Composite.

[a]Test–retest reliability coefficients for normative data found in test manuals.

[b]Statistically significant differences between Pearson's test–retest correlation coefficients of current sample and those reported in test manuals using Fisher r-to-z transformation (2-tailed significance, $p < .05$).

*Reliability averaged across age groups.

Table 4. Reliable change in children with unilateral cerebral palsy ($n = 44$) for neuropsychological measures

Measures	S_{diff}	RCI 90% CI	Minimal required for clinically meaningful change	Declined (%)	Improved (%)	Stable (%)
WISC-IV ($n = 43$)						
Digit Span Forwards	2.09	3.43	3 Scaled Scores	0.12	0.21	0.67
Digit Span Backwards	2.47	4.07	4 Scaled Scores	0.09	0.02	0.89
Digit Span Total	1.99	3.27	3 Scaled Scores	0.05	0.12	0.83
Coding	1.88	3.09	3 Scaled Scores	0.05	0.14	0.81
Symbol Search	2.20	3.63	4 Scaled Scores	0.02	0.05	0.93
D-KEFS TMT ($n = 44$)						
Number-Sequencing	3.17	5.21	5 Scaled Scores	0.07	0.07	0.86
Letter-Sequencing	3.12	5.14	5 Scaled Scores	0.05	0.16	0.79
Number-Letter Seq.	3.20	5.27	5 Scaled Scores	0.07	0.14	0.79
Motor Speed	2.54	4.18	4 Scaled Scores	0.07	0.07	0.86
D-KEFS Color-Word Test ($n = 44$)						
Word Reading	1.96	3.23	0.11	0.11	0.05	0.84
Color Naming	2.04	3.35	3 Scaled Scores	0.16	0.05	0.79
Inhibition	2.19	3.60	4 Scaled Scores	0.07	0.07	0.86
D-KEFS Tower Test						
Achievement	2.32	3.81	4 Scaled Scores	0.00	0.30	0.70
BRIEF ($n = 42$)						
Inhibit	10.9	17.92	18 T scores	0.05	0.05	0.90
Shift	9.57	15.75	16 T scores	0.05	0.05	0.90
Emotional Control	9.55	15.71	16 T scores	0.12	0.02	0.86
Initiate	10.4	17.06	17 T scores	0.00	0.02	0.98
WM	7.10	11.68	12 T scores	0.00	0.10	0.90
Plan/Organize	6.22	10.23	10 T scores	0.00	0.12	0.88
Org. of Materials	5.19	8.54	9 T scores	0.05	0.10	0.85
Monitor	6.08	10	10 T scores	0.02	0.07	0.91
BRI	8.54	14.05	14 T scores	0.07	0.05	0.88
MCI	7.23	11.90	12 T scores	0.00	0.10	0.90
GEC	5.85	9.63	10 T scores	0.02	0.07	0.91

S_{diff} = standard error of difference; RCI = Reliable Change Index; % = percentage; CI = confidence interval; WISC-IV = Wechsler Intelligence Scale for Children – Fourth Edition; D-KEFS = Delis-Kaplan Executive Function System; seq. = sequence; TMT = Trail Making Test; BRIEF = Behavior Rating Inventory of Executive Function; WM = working memory; Org. = Organization; BRI = Behavioral Rating Index; MCI = Metacognition Index; GEC = Global Executive Composite.

$p = .4$; Inhibition of the Color-Word Interference Test (D-KEFS) $t(43) = 2.12$, $p = .4$; Tower Achievement $t(43) = 5.60$, $p < .00$; and the Metacognition Index $t(41) = 2.30$, $p = .03$ of the BRIEF. The 90% confidence interval for calculating reliable change is presented in Table 4. These reliable change scores were applied to the original data in order to calculate the percentage of participants that would be classified as reliably improved, declined or stable. These results are also presented in Table 4.

DISCUSSION

The current study aimed to examine the test–retest stability of a range of EF measures in children with mild to moderate unilateral CP aged between 8 and 17 years. It was anticipated that all measures would remain stable over the 20-week retest period as the participants were not involved in an intervention program. Most of the measures used in the current study demonstrated excellent to fair Pearson's test–retest reliabilities ($r = 0.91$–0.74), confirming their stability over testing periods in this population. Test–retest reliabilities, however, were unacceptable for WISC-IV Digit Span Backwards, the Inhibition condition of the D-KEFS Color-Word Interference Test, and the Initiate score on the BRIEF. This indicates that these scores are both unreliable and unstable across administrations.

For most of the measures, test–retest reliabilities were not significantly different ($p > .05$) to those reported for TDC (see Table 3), however the test–retest reliabilities for WISC-IV Symbol Search; Letter Sequencing and Number Letter Sequencing of the D-KEFS TMT; Color Naming of the D-KEFS Color Word Interference Test; and the Monitor score of the BRIEF were statistically better in the current study compared to those reported in the test manuals. On the other hand, the test–retest reliability in the current sample for the Inhibition condition of the D-KEFS Color-Word Interference Test was statistically worse than those reported in the manual.

To account for the variability and error associated with repeated administrations of neuropsychological assessments, change scores based on the RCI method employed by Iverson and Green (2001) are provided for use for practitioners and researchers. These change scores apply 90% confidence intervals for estimating reliable change and account for practice effects. This is particularly important in this battery of neuropsychological tests, as many demonstrated a mean practice effect between the initial and second testing session. By rounding the values derived from the table, clinicians and researchers can determine the change in score required to detect clinically meaningful and significant change beyond measurement error and practice effects. By applying these change scores to the original data it is evident that the measures used in the current study are highly stable in children and adolescents with unilateral CP. Out of the participants in the study, 67–98% showed no improvement or decline between testing sessions.

As a guide for clinicians and researchers conducting repeated neuropsychological administrations in similar unilateral CP populations, the reliable change scores to detect clinically meaningful change are reported in Table 4. To understand the clinical relevance of these results, as an example, a child with unilateral CP who obtains a scaled score of 7 (Low Average range) on WISC-IV Coding in their initial testing session would need to improve to at least a scaled score of 10 (Average range) on their retest for a clinician and/or researcher to be 90% confident that the score reflects clinical and significant change beyond measurement error and practice effects. Likewise, for a clinician and/or researcher to be 90% confident that clinical and significant change has occurred between testing intervals, a child with unilateral CP must have a retest score that is more than or equal to 10 T-scores above their initial T-score for GEC, an improvement of one standard deviation.

An important implication of this study is that these change scores allow clinicians and/or researchers to evaluate cognitive and EF performance in children with unilateral

CP across repeated neuropsychological administrations. Previously, a lack of test–retest reliability data for this population meant that comparisons were made with data from TDC published in test manuals. This paper will allow clinicians and researchers to make more informed and evidence-based decisions about performance across administrations either from a developmental point of view or from a research perspective where interventions targeted at improving EF are implemented in this target population.

There are some potential limitations to this study. Although every attempt was made to ensure consistency during testing, testing results might not correctly represent true cognitive ability at time of testing due to confounding variables such as patient and environment factors. Another limitation is the relatively small sample size of the study and the wide age range of the participants. As a result, this study was not able to examine developmental differences within the sample. Due to the characteristics of the participants, results should not be generalized to children and adolescents with a more severe diagnosis of CP or bilateral CP. Future research should be aimed at developing RCI change scores for EF measures in these diagnostic groups given the extensive research that suggests EF difficulties. There is also considerable variability between the FSIQs of participants in this study. Rapport, Brines, Axelrod, and Theisen (1997) have demonstrated that practice effects seen between test administrations may be dependent on initial FSIQ scores, with individuals with higher IQ scores at their first assessment showing a greater improvement at subsequent assessments. This might potentially influence the RCI calculated for the participants in this study; however this is dependent on how homogeneous is each of the samples. Overall, it is more the case that low average FSIQ would affect the generalizability of the results, not the internal validity of the study. Future research should aim to establish whether differences in FSIQ influence RCI calculations and whether RCI change scores should be calculated based on level of intellectual functioning.

Lastly, there is still considerable debate regarding the most reliable RCI. The authors have chosen the RCI employed by Iverson and Green (2001) as it was deemed most appropriate and accounted for practice effects given the high level of mean practice effects seen in the data.

CONCLUSION

This study demonstrated excellent to fair test–retest stability for a range of neuropsychological tests in 8–17-year-old children with unilateral CP, with the exception of the WISC-IV Digit Span Backwards; Inhibition condition of the D-KEFS Color Word Interference Test; and the BRIEF Initiate score. These measures demonstrated inadequate Pearson's test–retest reliability ($r = 0.62$–0.69). These are not consistent with test–retest data published in the test manuals. Based on the current findings it is recommended that the Coding and Symbol Search of the WISC-IV; D-KEFS TMT; D-KEFS Tower Test; and the BRIEF are used to examine EF and neuropsychological performance over repeated administrations in children and adolescents with mild to moderate unilateral CP. These measures were shown to remain stable over the retest period for this population group and had fair to excellent test–retest reliabilities. Furthermore, it is recommended that the D-KEFS Color Word Interference test is not

used as a measure of EF attentional control, and the WISC-IV Digit Span Backwards is not used as a measure of EF cognitive flexibility in this population given the inadequate stability and reliability.

ACKNOWLEDGMENTS

REFERENCES

Alexander, M., & Stuss, D. (2000). Disorders of frontal lobe functioning. *Seminars in Neurology, 20*, 427–437.

Anderson, P. (2002). Assessment and development of executive function (EF) during childhood. *Child Neuropsychology, 8*, 71–82.

Anderson, V., Spencer-Smith, M., Coleman, L., Anderson, P., Williams, J., Greenham, M., … Jacobs, R. (2010). Children's executive functions: Are they poorer after early brain insult. *Neuropsychologia, 48*, 2041–2050.

Anderson, V., Spencer-Smith, M., & Wood, A. (2011). Do children really recover better? Neuro-behavioural plasticity after early brain insult. *Brain, 134*, 2197–2221.

Bax, M., Goldstein, M., Rosenbaum, P., Leviton, A., Paneth, N., Dan, B., … Damiano, D. (2005). Proposed definition and classification of cerebral palsy, April 2005. *Developmental Medicine & Child Neurology, 47*, 571–576.

Barr, W. (2002). Neuropsychological testing for assessment of treatment effects: Methodologic issues. *CNS Spectrums, 7*, 300–306.

Bodimeade, H. L., Whittingham, K., Lloyd, O., & Boyd, R. N. (2013). Executive function in children and adolescents with unilateral cerebral palsy. *Developmental Medicine & Child Neurology, 55*, 926–933.

Bottcher, L., Flachs, E. M., & Uldall, P. (2010). Attentional and executive impairments in children with spastic cerebral palsy. *Developmental Medicine & Child Neurology, 52*, e42–e47.

Boyd, R. N., Mitchell, L. E., James, S. T., Ziviani, J., Sakzewski, L., Smith, A., … Rasmussen, B. (2013). Move it to improve it (Mitii™): A randomized controlled trial of a novel web-based multi-modal. *BMJ Open, 3*, 1–20.

Burnett, A. C., Scratch, S. E., & Anderson, P. J. (2013). Executive function outcome in preterm adolescents. *Early Human Development, 89*, 215–220.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284–290.

Collette, F., Hogge, M., Salmon, E., & Van, d. L. M. (2006). Exploration of the neural substrates of executive functioning by functional neuroimaging. *Neuroscience, 139*, 209–221.

Crawford, J. R., Anderson, V., Rankin, P. M., & MacDonald, J. (2010). An index-based short-form of the WISC-IV with accompanying analysis of the reliability and abnormality of differences. *British Journal of Clinical Psychology, 49*, 235–258.

de Vet, H. C., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology, 59*, 1033–1039.

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

Della Sala, S., Gray, C., Spinnler, H., & Trivello, C. (1998). Frontal lobe functioning in man: The riddle revisited. *Archives of Clinical Neuropsychology, 13*, 663–682.

Duncan, J. (1995). Attention, intelligence, and frontal loves. In M. Gazzaniga (Ed.-in-Chief), *The cognitive neurosciences* (pp. 721–733). Cambridge, MA: MIT.

Edgin, J. O., Inder, T. E., Anderson, P. J., Hood, H. M., Clark, C. A. C., & Woodward, L. J. (2008). Executive function in preschool children born very preterm: Relationship with early white matter pathology. *Journal of the International Neuropsychological Society, 14*, 90–101.

Eliasson, A. C., Krumlinde-Sundholm, L., Rosblad, B., Beckung, E., Arner, M., Ohrvall, A. M., & Rosenbaum, P. (2006). The Manual Ability Classification System (MACS) for children with cerebral palsy: Scale development and evidence of validity and reliability. *Developmental Medicine & Child Neurology, 48*, 549–554. doi:10.1017/s0012162206001162

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

Goodman, R. (1999). The extended version of the strengths and difficulties questionnaire as a guide to psychiatric caseness and consequent burden. *Journal of Child Psychology, 40*, 791–799. doi:10.1111/1469-7610.00494

Guzzetta, A., Mercuri, E., & Cioni, G. (2001). Visual disorders in children with brain lesions: 2. Visual impairment associated with cerebral palsy. *European Journal of Paediatric Neurology, 5*, 115–119. doi:10.1053/ejpn.2001.0481

Himmelmann, K., & Uvebrant, P. (2014). The panorama of cerebral palsy in Sweden. XI. Changing patterns in the birth-year period 2003–2006. *Acta Paediatrica, 103*, 618–624.

Hinton-Bayre, A. D. (2010). Deriving reliable change statistics from test–retest normative data: Comparison of models and mathematical expressions [comparative study]. *Archives of Clinical Neuropsychology, 25*, 244–256.

Hinton-Bayre, A. D. (2012). Choice of reliable change model can alter decisions regarding neuropsychological impairment after sports-related concussion. *Clinical Journal of Sports Medicine, 22*, 105–108.

Iverson, G. L., & Green, P. (2001). Measuring improvement or decline on the WAIS-R in inpatient psychiatry. *Psychological Reports, 89*, 457–462.

Iverson, G. L., Lovell, M. R., & Collins, M. W. (2003). Interpreting change on ImPACT following sport concussion. *The Clinical Neuropsychologist, 17*, 460–467.

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.

Kennes, J., Rosenbaum, P., Hanna, S. E., Walter, S., Russell, D., Raina, P., ... Galuppi, B. (2007). Health status of school-aged children with cerebral palsy: Information from a population-based sample. *Developmental Medicine & Child Neurology, 44*, 240–247.

Krigger, K. W. (2006). Cerebral palsy: An overview. *American Family Physician, 73*, 91–100.

Lezak, M. D. (1995). *Neuropsychological assessment*. New York, NY: OUP.

Liu, U. M., Thawinchain, N., Palizano, R. J., & Valvano, J. (1998). The interrater reliability and stability of the Gross Motor Function Classification System. *Pediatric Physical Therapy, 10*, 174–175.

Odding, E., Roebroeck, M. E., & Stan, H. J. (2006). The epidemiology of cerebral palsy: Incidence, impairments, and risk factors. *Disability Rehabilitation, 28*, 183–191.

Palisano, R. J., Rosenbaum, P., Walter, S., Russell, D., Wood, E., … Galuppi, B. (2000). Validation of a model if gross motor function for children with cerebral palsy. *Physical Therapy, 80*, 974–985.

Pirila, S., van der Meere, J., Pentikainen, T., Ruusu-Niemi, P., Korpela, R., & Kilpinen, J. (2007). Language and motor speech skills in children with cerebral palsy. *Journal of Communication Disorders, 40*, 116–128.

Pirila, S., van der Meere, J. J., Rantanen, K., Jokiluoma, M., & Eriksson, K. (2011). Executive functions in youth with spastic cerebral palsy. *Journal of Child Neurology, 26*, 817–821.

Rosenbaum, P., Paneth, N., Leviton, A., Goldstein, M., & Bax, M. (2006). A report: The definition and classification of cerebral palsy April 2006. *Developmental Medicine & Child Neurology, 49*, 8–14.

Stanley, F., Blair, E., & Alberman, E. (2000). *Cerebral palsies: Epidemiology and causal pathways: Clinics in developmental medicine no. 151*. London: MacKeith Press.

Straub, K., & Obrzut, J. E. (2009). Effects of cerebral palsy on neuropsychological function. *Journal of Developmental and Physical Disability, 21*, 153–167.

Stuss, D., & Alexander, M. (2000). Executive functions and the frontal lobes: A conceptual view. *Psychologucal Research, 63*, 289–298.

The Australian Cerebral Palsy Register Group. (2013). *Report of the Australian Cerebral Palsy Register, birth years 1993–2006*. Canberra: Author.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children – Fourth* edition: *Technical and interpretative manual*. San Antonio, TX: The Psychological Corporation.

Whittingham, K., Bodimeade, H. L., Lloyd, O., & Boyd, R. N. (2014). Everyday psychological functioning in children with unilateral cerebral palsy: Does executive functioning play a role? *Developmental Medicine & Child Neurology, 56*, 572–579.

Wood, E., & Rosenbaum, P. (2000). The Gross Motor Function Classification System for cerebral palsy: A study of reliability and stability over time. *Developmental Medicine & Child Neurology, 42*, 292–296.