OXFORD
UNIVERSITY PRESS

Archives of Clinical Neuropsychology 32 (2017) 1001–1009

Archives of CLINICAL NEUROPSYCHOLOGY

Reliable Change on Memory Tests is Common in Healthy Children and Adolescents

Brian L. Brooks[1,2,3,*], James A. Holdnack[4], Grant L. Iverson[5]

[1]*Neurosciences Program (Brain Injury and Rehabilitation), Alberta Children's Hospital, Calgary, Alberta, Canada*
[2]*Departments of Paediatrics, Clinical Neurosciences, and Psychology, University of Calgary, Calgary, Alberta, Canada*
[3]*Alberta Children's Hospital Research Institute, University of Calgary, Calgary, Alberta, Canada*
[4]*Pearson Clinical Assessment, San Antonio, TX, USA*
[5]*Department of Physical Medicine and Rehabilitation, Harvard Medical School, Spaulding Rehabilitation Hospital & Mass General Hospital for Children Sport Concussion Program, Boston, MA, USA*

*Corresponding author at: Neurosciences Program, Alberta Children's Hospital, 2888 Shaganappi Trail NW, Calgary, Alberta T3B 6A8, Canada. Tel.: +1-403-955-2597; fax: +1-403-955-7501.
E-mail address: brian.brooks@ahs.ca (B.L. Brooks).

Editorial Decision 13 March 2017; Accepted 21 March 2017

Abstract

Objective: Neuropsychologists interpret a large number of scores in their assessments, including numerous retest scores to determine change over time. The rate at which healthy children and adolescents obtain reliably improved or declined memory scores when retested has yet to be explored. The purpose of this study was to illustrate the prevalence of reliable change scores on memory test batteries in healthy children and adolescents.
Methods: Participants were children and adolescents from test–retest samples from two published memory test batteries (ChAMP and CMS). Reliable change scores (RCI with 90% confidence interval and practice effects) were calculated for the indexes and subtests of each battery. Multivariate base rates involved considering all change scores simultaneously within each battery and calculating the frequencies of healthy children obtaining one or more reliably declined or one or more reliably improved scores.
Results: Across both memory batteries, one or more reliably changed index or subtest score was common; however, reliable change on three or more scores was uncommon (i.e., found in <5% of the samples). Base rates of change scores did not differ by parent education.
Conclusions: Having a single reliably changed score on retest is common when interpreting these memory batteries. Multivariate interpretation is necessary when determining cognitive decline and cognitive recovery. Further research is warranted with other measures, other samples, and different retest intervals.

Keywords: Test–retest; Psychometrics; RCI; Cognition; Multivariate

Introduction

Repeated neuropsychological testing is used to monitor children's cognitive development, modify academic accommodations for those with neurodevelopmental problems, evaluate treatment and rehabilitation efforts, or monitor recovery and outcome from an acquired brain injury. However, the rate at which children and adolescents obtain statistically meaningful changes in test scores, in the absence of a change in their clinical condition or cognitive development, is unknown. A small number of prior studies illustrated that it is common (i.e., prevalence rates ranging from 25% to 63%) for student athletes, adults, and older adults to obtain at least one statistically meaningful change in performance, improvement or decline, when multiple neuropsychological tests are administered on a second occasion (Brooks, Holdnack, & Iverson, 2016; Iverson, Lovell, & Collins, 2003; Woods et al., 2006). In a recent study examining the multivariate base rates of reliable change scores (i.e., when considering performance on all change scores simultaneously), having a statistically meaningful change in one or more memory scores on retest was found in 63% of healthy adults and 52% of healthy older adults (Brooks et al., 2016).

doi:10.1093/arclin/acx028 Advance Access publication on 5 April 2017

Using this multivariate approach to interpreting change scores, having two or more memory scores reliably change on retest was considered uncommon in the full retest sample and those with higher intellectual abilities were more likely to experience change on retest. The same study found that there were not differences based on sex, education, or race. Brooks et al. suggested that interpretation of an isolated reliable change score, when multiple scores were administered, may lead to an erroneous interpretation of decline or improvement. More research is needed, and in particular with children, when considering the implications of these findings for clinical practice and research.

The purpose of this study was to examine the prevalence of reliable change scores on memory testing in healthy children and adolescents. The reliable change methodology, which is widely used in clinical neuropsychology (Barr & McCrea, 2001; Chelune, Naugle, Luders, Sedlak, & Awad, 1993; Heaton et al., 2001; Hinton-Bayre, Geffen, Geffen, McFarland, & Friis, 1999; Iverson et al., 2003; Iverson, 1998, 1999, 2001; Temkin, Heaton, Grant, & Dikmen, 1999), was employed to estimate statistically reliable improvement or decline in performance. It was hypothesized that having at least one large improvement or decline on memory testing would be common when children are re-administered a battery of memory tests.

Methods

Participants

Participants for this study were from two test–retest samples that were included in the standardization process for their respective published memory batteries. These measures include the: (a) Child and Adolescent Memory Profile (ChAMP; Sherman & Brooks, 2015) and (b) Children's Memory Scale (CMS; Cohen, 1997).

Measures

For the ChAMP, there were two retest samples included in the standardization process; one shorter-interval retest sample and one longer-interval retest sample. For this study, the longer-interval sample (i.e., "reliable change sample") and the subsequent reliable change scores published in Table J.1 (Sherman & Brooks, 2015) were used. This retest sample was selected because of the longer retest interval and the use of their data for establishing reliable change scores in the professional manual. The ChAMP reliable change sample was selected from the larger standardization sample and those subjects were re-administered the battery a second time (see page 45, Sherman and Brooks, 2015). For all ages, all five ChAMP index scores (Verbal Memory, Visual Memory, Immediate Memory, Delayed Memory, and Total Memory) and all 10 ChAMP subtest age-adjusted standard scores (Lists, Lists Delayed, Lists Recognition, Instructions, Instructions Delayed, Instructions Recognition, Objects, Objects Delayed, Places, and Places Delayed) were considered.

For the CMS, the retest sample consisted of 125 youth between 5 and 16 years of age who were included in the standardization sample and were administered the battery on two occasions (see p. 99, Cohen, 1997). All eight index scores (Visual Immediate, Visual Delayed, Verbal Immediate, Verbal Delayed, General Memory, Attention/Concentration, Learning, and Delayed Recognition) were considered for this study. A total of 10 CMS subtest scores were considered that included sufficient data for all ages (i.e., not all CMS tests were administered a second time, so only those tests with all data were included in the multivariate analyses). There were six primary memory subtest age-adjusted standard scores, including: Dot Locations Learning, Dot Locations Long Delay, Faces Immediate, Faces Delayed, Word Pairs Long Delay, and Word Pairs Delayed Recognition. Stories Immediate and Stories Delayed were not given to all ages. Instead, Word List Learning and Word List Free Recall Delay were included in these analyses as substitutes. Numbers Forward and Numbers Backward were included as attention/concentration scores.

Analyses

Computation of the reliable change index (RCI) was done through a modification of the method proposed by Jacobson and Truax (1991), which has been used previously to present change scores that can be interpreted clinically (for example, see Iverson, 2012). The reliable change methodology allows the clinician to estimate measurement error surrounding test–retest difference scores using information commonly presented in the test manuals (i.e., standard deviations at time 1 and time 2, and test–retest correlations). All analyses were conducted with age-adjusted standard scores and not raw scores. For this study, we used RCI with a 90% confidence interval and included practice effects. Practice effects were defined as the difference in mean group scores between Time 2 and Time 1, which was subsequently added to the RCI and is a recommended process when interpreting repeated memory testing (Duff, 2012). The steps for calculating reliable change scores are provided below and Table 1 shows the values used in this study. For the ChAMP, reliable change scores were already presented in the

Table 1. Reliable change scores used for ChAMP and CMS

Subtest and index names	SD_1	SD_2	r_{12}	SEM_1	SEM_2	SE_{diff}	Mean PE	RCI (90% CI with PE)	
								Declined	Improved
ChAMP									
Lists	2.8	2.7	.68	1.6	1.5	2.2	1.9	−1	+5
Lists Delayed	3.2	2.8	.49	2.3	2.0	3.0	2.0	−3	+7
Objects	2.9	3.4	.56	1.9	2.3	3.0	2.1	−3	+7
Objects Delayed	3.0	2.8	.54	2.0	1.9	2.8	1.5	−3	+5
Instructions	2.8	3.1	.52	1.9	2.1	2.9	0.6	−4	+6
Instructions Delayed	3.1	3.5	.52	2.1	2.4	3.2	0.3	−6	+6
Places	2.6	2.6	.35	2.1	2.1	3.0	1.9	−3	+7
Places Delayed	2.8	3.0	.27	2.4	2.6	3.5	1.9	−4	+8
Lists Recognition	3.3	2.8	.53	2.3	1.9	3.0	1.0	−4	+6
Instructions Recognition	3.6	3.1	.40	2.8	2.4	3.7	−0.1	−6	+6
Verbal Memory Index	13.4	14.5	.60	8.5	9.2	12.5	7.1	−12	+26
Visual Memory Index	12.7	13.4	.53	8.7	9.2	12.7	10.4	−9	+29
Immediate Memory Index	11.5	13.1	.69	6.4	7.3	9.7	10.5	−3	+23
Delayed Memory Index	12.4	13.5	.56	8.2	9.0	12.2	9.0	−9	+27
Total Memory Index	11.7	13.5	.68	6.6	7.6	10.1	10.2	−4	+24
CMS									
Dot Locations Learning	2.5	2.8	.58	1.6	1.8	2.5	0.7	−4	+5
Dot Locations Long Delay	2.5	2.4	.53	1.7	1.7	2.4	0.5	−4	+5
Faces Immediate	2.8	3.1	.59	1.8	2.0	2.6	2.6	−2	+7
Faces Delayed	3.1	3.2	.50	2.2	2.2	3.1	2.2	−3	+8
Word Pairs Long Delay	3.3	3.5	.69	1.8	2.0	2.7	1.8	−3	+7
Word Pairs Delayed Recognition	3.1	3.1	.45	2.3	2.3	3.3	0.2	−6	+6
Word List Learning	3.1	3.2	.66	1.8	1.9	2.6	1.7	−3	+6
Word List Free Recall	3.0	3.1	.58	2.0	2.0	2.8	1.3	−4	+6
Numbers Forward	2.7	2.9	.72	1.5	1.5	2.1	0.8	−3	+5
Numbers Backward	3.1	3.3	.62	1.9	2.1	2.8	0.0	−5	+5
Visual Immediate Index	12.3	14.0	.59	7.9	9.0	12.0	11.1	−9	+31
Visual Delayed Index	12.9	13.6	.57	8.5	8.9	12.3	8.6	−12	+29
Verbal Immediate Index	16.8	18.0	.85	6.6	7.1	9.7	12.0	−4	+28
Verbal Delayed Index	16.5	17.1	.74	8.4	8.7	12.1	11.4	−9	+32
General Memory Index	16.8	16.0	.85	6.5	6.2	9.0	16.3	+2	+31
Attention/Concentration Index	15.4	15.7	.86	5.9	6.0	8.4	3.9	−10	+18
Learning Index	14.1	14.9	.73	7.3	7.7	10.6	9.4	−8	+27
Delayed Recognition Index	14.8	16.8	.56	9.8	11.1	14.8	3.9	−21	+28

Note: *SD* = standard deviation. r_{12} = retest correlation. SEM = standard error of measurement. SE_{diff} = standard error of the difference. PE = practice effect (mean time 2 – mean time 1). RCI = reliable change index. ChAMP = Child and Adolescent Memory Profile™ (Sherman & Brooks, 2015). CMS = Children's Memory Scale (Cohen, 1997).

test manual (see Table J.1) so they did not need to be calculated again; however, they were calculated by the publisher using this same methodology.

(1) $SEM_1 = SD\sqrt{1 - r_{12}}$ Standard deviation from time 1 multiplied by the square root of 1 minus the test–retest coefficient.
(2) $SEM_2 = SD\sqrt{1 - r_{12}}$ Standard deviation from time 2 multiplied by the square root of 1 minus the test–retest coefficient.
(3) $SE_{diff} = \sqrt{SEM_1^2 + SEM_2^2}$ Standard error of the difference is the square root of the sum of the squared SEMs for each testing occasion.
(4) RCI 90% CI = $SE_{diff} \times 1.64$
(5) RCI 90% CI accounting for practice effects:
    (a) Reliable improvement = Positive RCI (90% CI) *plus* practice effect
    (b) Reliable decline = Negative RCI (90% CI) *plus* practice effect

Frequency distributions for reliable change scores were computed for both batteries, including the prevalence of 0, 1, 2, etc. reliable change scores when all change scores were considered simultaneously (multivariate interpretation). This was done for (a) any reliable change (either improvement *or* decline), (b) reliable improvement only, and (c) reliable decline only. Frequency distributions were further stratified by level of parent education (≤12 years and >12 years) because of research on prior multivariate analyses showing effects based on level of parent education (Brooks, 2010, 2011). Analyses to determine whether frequencies differed across groups were done using chi-square analyses. Due to the large number of analyses and potential for Type I errors, alpha was set at $p < .0027$ (.05/18 comparisons).

Results

Descriptions of the test–retest samples are presented in Table 2. Mean retest intervals for the pediatric memory batteries ranged from 2 months on the CMS to just over 3 months on the ChAMP. Samples were roughly evenly split between sexes and have a higher proportion of Caucasians. About two-thirds of the parents in the ChAMP and CMS retest samples had more than high school completed. The frequencies of reliable change scores, when considering each score on its own (univariate analyses), are presented at the bottom of Table 2. The mean percentages of healthy youth who obtained a reliably different score on retest were approximately at the expected 10% range for both indexes and subtests, and were generally symmetrical for declined (approximately 5%) and improved (approximately 5%).

The multivariate frequencies of reliable change scores on the ChAMP are presented in Table 3. Overall in the sample, 36.7% had one or more reliably different index scores in either direction (includes improved and declined together). There were no statistically significant differences in the frequencies of whether participants in the retest sample had declined or improved on their ChAMP index scores; 18.4% had one or more reliably declined index scores and 18.4% had one or more reliably improved index scores [$\chi^2(1) = 0.0$, $p = 1.0$]. The ChAMP Total Memory Index score reliably declined most often (33.3% of the time) compared to all other ChAMP Indexes [each declined 16.7% of the time; $\chi^2(1) = 1.33$, $p = .25$]. The ChAMP Total Memory Index reliably improved most often (28.6% of the time) compared to ChAMP Verbal Memory and Visual Memory indexes that reliably improved the least often [each 14.3% of the time; $\chi^2(1) = 1.27$, $p = .26$]. There were no significant differences in the base rates of reliably declined [$\chi^2(1) = 0.01$, $p = .92$] or reliably improved [$\chi^2(1) = 0.01$, $p = .92$] index scores when grouping by parent education (≤12 years vs. >12 years).

Table 2. Test–retest sample characteristics for the pediatric memory batteries

Demographics	ChAMP		CMS	
N	49		125	
Mean age, years (*SD*)	13.0 (4.3); Range = 6–21		9.8 (3.2); Range = 5–16	
Sex (% Female)	55		48	
Ethnicity (% Caucasian)				
Caucasian	61.2		78.0	
African American	8.2		11.0	
Hispanic	20.4		10.0	
Other	10.2		1.0	
Parent education (% with >12 years of education)	65.3		65.6	
Mean retest interval, weeks (range)	14.7 (8.7–20.1)		8.5 (0–12.0)	
Univariate frequencies of reliable change scores[a]	Decline	Improve	Decline	Improve
Indexes				
Mean percent of sample showing reliable change	7.3	8.6	5.6	4.3
SD	2.7	2.7	1.6	1.8
Range	6.1–12.2	6.1–12.2	3.2–8.1	2.4–8.0
Subtests				
Mean percent of sample showing reliable change	6.5	7.2	5.1	5.9
SD	3.2	3.2	3.0	2.5
Range	2.0–10.2	2–14.3	0–8.8	3.2–9.6

Note: *SD* = standard deviation; %= percentage.

ChAMP: Reproduced by special permission of the Publisher, Psychological Assessment Resources, Inc. 16204 North Florida Avenue, Lutz, Florida 33549, from the *Child and Adolescent Memory Profile* by Elisabeth M.S. Sherman, Ph.D. and Brian L. Brooks, Ph.D., Copyright 2015 by PAR, Inc. Further reproduction is prohibited without permission from PAR, Inc.

CMS: Standardization data from the *Children's Memory Scale (CMS)*. Copyright © 1997 NCS Pearson, Inc. Used with permission. All rights reserved.

[a]Univariate frequencies of reliable change scores represent the percent of the sample identified as reliably declined or reliably improved if considering each change score in isolation.

Table 3. Cumulative percentages of the ChAMP sample with reliably different index or subtest scores on retest when considering all change scores simultaneously

	Cumulative percentages with number of reliably *declined* ChAMP scores	Cumulative percentages with number of reliably *improved* ChAMP scores
Number of index scores		
Zero	81.6	81.6
1 or more	18.4	18.4
2 or more	10.2	12.2
3 or more	6.1	10.2
4 or more	2.0	2.0
Number of subtest scores		
Zero	57.1	59.2
1 or more	42.9	40.8
2 or more	20.4	22.4
3 or more	2.0	4.1
4 or more	—	4.1

Note: All five ChAMP index scores (Verbal Memory, Visual Memory, Immediate Memory, Delayed Memory, and Total Memory) and all 10 ChAMP subtest scores (Lists, Lists Delayed, Lists Recognition, Instructions, Instructions Delayed, Instructions Recognition, Objects, Objects Delayed, Places, and Places Delayed) were considered.

For the ChAMP subtests, 77.6% of the sample had one or more reliably different scores in either direction. There were no statistically significant differences in the frequencies of whether participants in the retest sample had declined or improved on their ChAMP subtest scores; 42.9% had one or more reliably declined subtest scores and 40.8% had one or more reliably improved subtest scores [$\chi^2(1) = .04$, $p = 84$]. ChAMP Lists and Objects subtests reliably declined most often (each 15.6% of the time) compared to ChAMP Lists Recognition and Places Delayed that reliably declined the least often [each 3.1% of the time; $\chi^2(1) = 2.94$, $p = .09$]. The ChAMP Objects Delayed subtest reliably improved most often (20.0% of the time) compared to the ChAMP Instructions Delayed subtest that reliably improved the least often [2.9% of the time; $\chi^2(1) = 5.08$, $p = .02$]. There were no significant differences in the base rates of reliably declined [$\chi^2(1) = 2.71$, $p = .10$] or reliably improved [$\chi^2(1) = 0.33$, $p = .57$] subtest scores when considering parent education (≤12 years vs. > 12 years).

Multivariate base rates of reliable change scores for the CMS are presented in Table 4. One or more reliably changed index scores, in either direction, were found in 45.6% of the CMS retest sample. There were no statistically significant differences in the frequencies of whether participants in the retest sample had declined or improved on their CMS index scores; 26.2% had one or more reliably declined index scores and 28.2% had one or more reliably improved index scores [$\chi^2(1) = 0.10$, $p = .75$]. The CMS Verbal Immediate Memory Index reliably declined most often (17.9% of the time) compared to CMS Visual Immediate Memory and CMS Visual Delayed Memory Indexes that reliably declined the least often [each 7.7% of the time; $\chi^2(1) = 1.84$, $p = .18$]. The CMS Attention/Concentration Index reliably improved most often (24.4% of the time) compared to CMS Verbal Delayed, General Memory, and Delayed Recognition indexes that reliably improved the least often [each 7.3% of the time; $\chi^2(1) = 3.48$, $p = .03$]. There were no significant differences in the base rates of reliably declined [$\chi^2(1) = 3.52$, $p = .06$] or reliably improved [$\chi^2(1) = 0.46$, $p = 0.50$] index scores when considering parent education that is either ≤12 years or >12 years.

For the CMS subtests, 72.1% of the sample had one or more reliably different scores in either direction. There were no statistically significant differences in the frequencies of whether participants in the retest sample had declined or improved on their CMS subtest scores; 40.2% had one or more reliably declined subtest scores and 48.4% had one or more reliably improved subtest scores [$\chi^2(1) = 1.62$, $p = .20$]. The CMS Numbers Forward subtest reliably declined most often (18.0% of the time) compared to CMS Numbers Backward that reliably declined the least often [0.0% of the time; $\chi^2(1) = 12.09$, $p = .001$]. The CMS Word List Free Recall subtest reliably improved most often (16.7% of the time) compared to CMS Faces Delayed, Word List Learning, and Word Pair Delay subtests that reliably improved the least often [each 5.6% of the time; $\chi^2(1) = 4.40$, $p = .03$]. There were no significant differences in the base rates of reliably declined [$\chi^2(1) = 1.66$, $p = .20$] or reliably improved [$\chi^2(1) = 0.45$, $p = .50$] subtest scores when grouping by parent education (≤12 years vs. >12 years).

Discussion

It is common for healthy children and adolescents to have at least one statistically meaningful change in performance on memory testing during repeat testing. One-third of children retested on the ChAMP and nearly one-half of children retested

Table 4. Cumulative percentages of the CMS sample with reliably different index or subtest scores on retest when considering all change scores simultaneously

	Cumulative percentages with number of reliably *declined* CMS scores	Cumulative percentages with number of reliably *improved* CMS scores
Number of index scores		
Zero	73.8	71.8
1 or more	26.2	28.2
2 or more	6.8	6.8
3 or more	3.9	1.9
4 or more	1.0	—
Number of subtest scores		
Zero	59.8	51.6
1 or more	40.2	48.4
2 or more	8.2	8.2
3 or more	0.8	2.5
4 or more	0.8	—

Note: For all ages, all eight CMS index scores (Visual Immediate, Visual Delayed, Verbal Immediate, Verbal Delayed, General Memory, Attention/Concentration, Learning, and Delayed Recognition) were considered. A total of 10 CMS subtest scores were considered. There were six primary memory subtest scores, including Dot Locations Learning, Dot Locations Long Delay, Faces Immediate, Faces Delayed, Word Pairs Long Delay, and Word Pairs Delayed Recognition. Stories Immediate and Stories Delayed were not given to all ages. Instead, Word List Learning and Word List Free Recall Delay were included in these analyses as substitutes. Numbers Forward and Numbers Backward were included as well for the attention/concentration scores.

on the CMS show at least one statistically meaningful change on an index score (in either direction). When considering the subtest scores, one or more reliably changed scores are found in nearly three-quarters of children on both the ChAMP and CMS (in either direction). The changes in performance occur in both directions (worsening and improvement), they are more likely to occur on subtests than on index scores. It is important to remember that there are more subtest scores being considered and the retest reliabilities are generally lower than for index scores, both of which could change the rates of reliable change in a multivariate base rate analysis. The high rate of reliable changes on repeat memory testing is likely surprising to many clinicians and researchers. Therefore, clinicians and researchers should know that when considering 6–10 memory test scores simultaneously, it is common for children and adolescents to have at least one statistically meaningful change in performance on retesting. In contrast, it is uncommon to obtain 2 or 3 reliably changed memory scores. Although this study specifically considered memory batteries, it is expected that similar results will be found for other types of cognitive batteries based on the larger multivariate base rates literature supporting that these types of psychometric findings are not battery specific (see Binder, Iverson, & Brooks, 2009 for a review).

If one considers an entire battery of neuropsychological tests being repeated (beyond just a memory battery), it is apparent that children will obtain more than one statistically meaningful change in performance. Three prior studies have reported that having at least one statistically meaningful change in performance on retesting is common in young athletes, adults, and older adults (Brooks et al., 2016; Iverson et al., 2003; Woods et al., 2006). When retesting healthy young athletes on average 6 days later using a computerized battery of tests, Iverson et al. (2003) found that 43% had one or more reliably changed scores (using 80% confidence interval); it was "uncommon" (3.6%) to have two or more of the five composite scores reliably change on retest. Woods et al. (2006) considered a multivariate approach to interpreting change on a broad neuropsychological battery when retesting was done nearly one year later. In their adult cohort, it was "uncommon" (<5%) to have six or more test scores (out of 14) reliably change on retest. Brooks et al. (2016) reported similar results for three different memory batteries when retesting occurred between 1 and 6 months; in most analyses it was uncommon (<5%) to have 3 or more memory scores reliably change on retest. Clearly, these multivariate base rates of change scores have important implications for clinical practice and research. Our perspective is that knowing these psychometric principles improves interpretation of test results.

The findings from this study are related to, and extend, the larger literature on multivariate interpretation of test scores from a neuropsychological assessment. Previous studies illustrate that it is common for healthy children, adults, and older adults to obtain one or more low scores when administered a battery of neuropsychological tests (Brooks & Iverson, 2010; Brooks, Holdnack, & Iverson, 2011; Brooks, Iverson, & Holdnack, 2013; Brooks, Iverson, & White, 2009; Crawford, Garthwaite, & Gault, 2007; Iverson, Brooks, White, & Stern, 2008). Obtaining low scores within specific cognitive domains, such as memory (Brooks, Iverson, Sherman, & Holdnack, 2009) and executive functioning (Brooks et al., 2013), also occurs much more frequently in children than would be expected or predicted from normative percentile ranks. Moreover, it is common to have at least one significant index score discrepancy when several discrepancies are considered simultaneously

(Crawford et al., 2007). When coupled with the existing studies showing that it is common to have scores reliably change on retest (Brooks et al., 2016; Iverson et al., 2003; Woods et al., 2006), it is clear that multivariate interpretation of test performance is necessary to limit over-interpretation of neuropsychological data.

The sample sizes used in this study were fairly small and limiting from the perspectives of generalizability and additional sub-analyses. Related to this, the retest interval was relatively short (2 months on the CMS and 3 months on the ChAMP) and may not approximate retest periods for some clinical circumstances. This study used RCI as the methodology for interpreting change scores, so the results may differ if a different method of interpreting change scores is employed (e.g., regression). However, RCI was chosen because it can be easily and readily calculated by clinicians and researchers for any score using the test–retest information presented in technical manuals. The reader should note that there is some circularity to the analyses in this study. The retest samples were used to derive the reliable change scores, and then used to evaluate the multivariate base rates of these reliable change scores. This is the same methodology used for many past studies involving base rates of low scores in normative samples. Having independent samples to validate the results in future research will be necessary.

An important issue to consider is the role of practice effects. For this study, we corrected the confidence interval for change scores by the average practice effect for each sample, and those practice effects were large. The practice effects were large in part, at least, because the retest interval was relatively brief (e.g., 8 weeks for the CMS and 15 weeks for the ChAMP) and the samples were youth with no known clinical conditions. Practice effects over greater retest intervals (e.g., years) or in clinical populations might be smaller. As such, it is possible that these findings may not always translate well to diverse clinical and research situations and settings—especially if practice effects differ substantially in other groups. Including practice effects when retesting memory abilities has been advocated in the literature (e.g., Duff, 2012). For the present data, despite including the practice effects and increasing the amount of change needed to deem a reliable improvement, there was still symmetry for the proportions of participants whose scores improved or declined. There will be situations in which clinical judgment should be used in combination with data presented in Table 1. For example, a clinician could choose to multiply the SE_{diff} by a different confidence interval (e.g., 70%, 80%, or 95%) and not correct for practice effects when interpreting retest performance in a clinical patient. If the confidence interval is lowered, to increase sensitivity to change, the clinician should appreciate that the prevalence of change scores across a battery of tests, in healthy children, will increase.

Concentrated efforts are needed to translate these findings into clinical practice. Base rates of test–retest change scores might be influenced by a diverse range of factors, including situational factors, measurement error, practice effects, initial level of performance (e.g., high or low), regression to the mean, or motivational factors (see Iverson, 2012 for a discussion of these issues). In clinical practice and research, several of these factors might be relevant in specific individual cases. The present study illustrates that having at least one statistically meaningful change in performance on retesting, within a cognitive domain such as memory, is common in healthy children and adolescents; in contrast, it is uncommon (<5%) to have three or more reliably changed scores. These results are consistent with numerous prior studies suggesting that base rates of "abnormal" or "clinically significant" findings increase when multiple scores are interpreted. This study is considered a starting point for understanding multivariate base rates of change scores, therefore, it will be important to replicate these findings in other samples of healthy youth who are independent from the derivation of retest variability data. It will also be important to extend the findings into samples with longer retest intervals and those with medical, psychological, and/or neurological diagnoses. To improve research and clinical practice, we need to better understand factors that influence improvement or worsening in cognitive test performance that are separate from a real developmental change or a change associated with a clinical condition.

Funding

No grant funding was provided for this study. Salary support for this research was provided to Brian Brooks by the Neurosciences program at the Alberta Children's Hospital (Conny Betuzzi) and the Canadian Institutes of Health Research (CIHR) Embedded Clinician Researcher Salary Award. Grant Iverson acknowledges support from the Mooney-Reed Charitable Foundation, the Heinz Family Foundation, ImPACT Applications, Inc., and the TBI Endpoints Development Initiative (i.e., a grant entitled Development and Validation of a Cognition Endpoint for Traumatic Brain Injury Clinical Trials).

Conflict of Interest

Brian Brooks receives royalties for tests published by Psychological Assessment Resources, Inc. [Child and Adolescent Memory Profile (ChAMP, Sherman and Brooks, 2015), Memory Validity Profile (MVP, Sherman and Brooks, 2015), and

Multidimensional Everyday Memory Ratings for Youth (MEMRY, Sherman and Brooks, 2017)]. He also receives royalties from a book that is cited in this manuscript (Pediatric Forensic Neuropsychology; Sherman & Brooks, 2012). James Holdnack is Senior Scientist with Pearson Assessment, which is the publisher for the Children's Memory Scale (CMS; Cohen, 1997). Grant Iverson has been reimbursed by the government, professional scientific bodies, and commercial organizations for discussing or presenting research at meetings, scientific conferences, and symposiums. He has a clinical practice in forensic neuropsychology involving individuals who have sustained mild TBIs. He has received honorariums for serving on research panels that provide scientific peer review of programs. He is a co-investigator, collaborator, or consultant on grants relating to mild TBI funded by several organizations. He has received grant funding from pharmaceutical companies to do psychometric research using neuropsychological tests. He has received research support from neuropsychological test publishing companies in the past, such as PAR, Inc., CNS Vital Signs, and ImPACT Applications, Inc. He receives royalties from books in neuropsychology and one neuropsychological test (Psychological Assessment Resources, Inc.).

Acknowledgements

The authors thank Psychological Assessment Resources Inc. and NCS Pearson Inc. for graciously providing the data necessary for this study. Thanks to Amy Kovacs BA and Heddy Clark Ph.D. for help with running analyses on the ChAMP data.

References

Barr, W. B., & McCrea, M. (2001). Sensitivity and specificity of standardized neurocognitive testing immediately following sports concussion. *Journal of the International Neuropsychological Society*, *7*, 693–702.

Binder, L. M., Iverson, G. L., & Brooks, B. L. (2009). To err is human: "Abnormal" neuropsychological scores and variability are common in healthy adults. *Archives of Clinical Neuropsychology*, *24*, 31–46.

Brooks, B. L. (2010). Seeing the forest for the trees: Prevalence of low scores on the Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV). *Psychological Assessment*, *22*, 650–656.

Brooks, B. L. (2011). A study of low scores in Canadian children and adolescents on the Wechsler Intelligence Scale For Children, Fourth Edition (WISC-IV). *Child Neuropsychology*, *17*, 281–289.

Brooks, B. L., Holdnack, J. A., & Iverson, G. L. (2011). Advanced clinical interpretation of the WAIS-IV and WMS-IV: Prevalence of low scores varies by level of intelligence and years of education. *Assessment*, *18*, 156–167.

Brooks, B. L., Holdnack, J. A., & Iverson, G. L. (2016). To change is human: "Abnormal" reliable change memory scores are common in healthy adults and older adults. *Archives of Clinical Neuropsychology*. DOI:10.1093/arclin/acw079.

Brooks, B. L., & Iverson, G. L. (2010). Comparing actual to estimated base rates of 'abnormal' scores on neuropsychological test batteries: Implications for interpretation. *Archives of Clinical Neuropsychology*, *25*, 14–21.

Brooks, B. L., Iverson, G. L., & Holdnack, J. A. (2013). Understanding multivariate base rates. In Holdnack J. A., Drozdick L., Weiss L. G., & Iverson G. L. (Eds.), *WAIS-IV, WMS-IV, & ACS: Clinical use and interpretation* (pp. 75–102). New York: Elsevier.

Brooks, B. L., Iverson, G. L., Koushik, N. S., Mazur-Mosiewicz, A., Horton, , A. M.Jr., & Reynolds, C. R. (2013). Prevalence of low scores in children and adolescents on the test of verbal conceptualization and fluency. *Applied Neuropsychology: Child*, *2*, 70–77.

Brooks, B. L., Iverson, G. L., Sherman, E. M., & Holdnack, J. A. (2009). Healthy children and adolescents obtain some low scores across a battery of memory tests. *Journal of the International Neuropsychological Society*, *15*, 613–617.

Brooks, B. L., Iverson, G. L., & White, T. (2009). Advanced interpretation of the Neuropsychological Assessment Battery (NAB) with older adults: Base rate analyses, discrepancy scores, and interpreting change. *Archives of Clinical Neuropsychology*, *24*, 647–657.

Chelune, G. J., Naugle, R. I., Luders, H., Sedlak, J., & Awad, I. A. (1993). Individual change after epilepsy surgery: Practice effects and base-rate information. *Neuropsychology*, *7*, 41–52.

Cohen, M. J. (1997). *Children's Memory Scale*. San Antonio, TX: The Psychological Corporation.

Crawford, J. R., Garthwaite, P. H., & Gault, C. B. (2007). Estimating the percentage of the population with abnormally low scores (or abnormally large score differences) on standardized neuropsychological test batteries: A generic method with applications. *Neuropsychology*, *21*, 419–430.

Duff, K. (2012). Evidence-based indicators of neuropsychological change in the individual patient: Relevant concepts and methods. *Archives of Clinical Neuropsychology*, *27*, 248–261.

Heaton, R. K., Temkin, N. R., Dikmen, S. S., Avitable, N., Taylor, M. J., & Marcotte, T. D. (2001). Detecting change: A comparison of three neuropsychological methods using normal and clinical samples. *Archives of Clinical Neuropsychology*, *16*, 75–91.

Hinton-Bayre, A. D., Geffen, G. M., Geffen, L. B., McFarland, K. A., & Friis, P. (1999). Concussion in contact sports: Reliable change indices of impairment and recovery. *Journal of Clinical and Experimental Neuropsychology*, *21*, 70–86.

Iverson, G. L. (1998). Interpretation of Mini-Mental State Examination scores in community-dwelling elderly and geriatric neuropsychiatry patients. *International Journal of Geriatric Psychiatry*, *13*, 661–666.

Iverson, G. L. (1999). Interpreting change on the WAIS-III/WMS-III in persons with traumatic brain injuries. *Journal of Cognitive Rehabilitation*, *July/August*, 16–20.

Iverson, G. L. (2001). Interpreting change on the WAIS-III/WMS-III in clinical samples. *Archives of Clinical Neuropsychology*, *16*, 183–191.

Iverson, G. L. (2012). Interpreting change on repeated neuropsychological assessments of children. In Sherman E. M. S., & Brooks B. L. (Eds.), *Pediatric forensic neuropsychology* (pp. 89–112). New York: Oxford University Press.

Iverson, G. L., Brooks, B. L., White, T., & Stern, R. A. (2008). Neuropsychological Assessment Battery (NAB): Introduction and advanced interpretation. In Horton A. M., & Wedding D. (Eds.), *The neuropsychology handbook* (3rd ed., pp. 279–343). New York: Springer Publishing Inc.

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.

Sherman, E. M. S., & Brooks, B. L. (2015). *Child and Adolescent Memory Profile (ChAMP)*. Lutz, FL: Psychological Assessment Resources, Inc.

Temkin, N. R., Heaton, R. K., Grant, I., & Dikmen, S. S. (1999). Detecting significant change in neuropsychological test performance: A comparison of four models. *Journal of the International Neuropsychological Society*, *5*, 357–369.

Woods, S. P., Childers, M., Ellis, R. J., Guaman, S., Grant, I., & Heaton, R. K. (2006). A battery approach for measuring neuropsychological change. *Archives of Clinical Neuropsychology*, *21*, 83–89.