To Change is Human: "Abnormal" Reliable Change Memory Scores are Common in Healthy Adults and Older Adults

Brian L. Brooks[1,2,3,*], James A. Holdnack[4], Grant L. Iverson[5,6,7,8]

[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, Boston, MA, USA
[6]Spaulding Rehabilitation Hospital, Boston, MA, USA
[7]MassGeneral Hospital for Children Sport Concussion Program, Boston, MA, USA
[8]Home Base, A Red Sox Foundation and Massachusetts General Hospital Program, Boston, MA, USA

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

Abstract

Objective: The rate at which people obtain reliably improved or declined cognitive test scores when retested, in the absence of a change in clinical condition, is largely unknown. The purpose of this study was to illustrate the prevalence of statistically reliable change scores on memory test batteries in healthy adults and older adults.

Method: Participants included three adult and older adult test-retest samples from memory test batteries. Reliable change scores (reliable change index with 90% confidence interval and practice effects) were calculated for the indexes and subtests of each battery. Multivariate analyses involved calculating the frequencies of healthy people obtaining one or more reliably declined or one or more reliably improved scores when considering all change scores simultaneously within each battery.

Results: Across all batteries, having one or more reliably changed index or subtest score on retest was common. With most batteries, having two or more reliably changed scores was uncommon. Those with higher intellectual abilities were more likely to have a change on retest; however, no significant differences in base rates were found based on education level, sex, or ethnic minority status. Those older adults who did not have any low memory scores were more likely to improve than decline on retest.

Conclusions: Having a single reliably changed score on retest is common when interpreting a battery of memory measures. This has implications for determining cognitive decline and cognitive recovery, suggesting that multivariate interpretation is necessary.

Keywords: Test-retest; Psychometrics; Reliable change index; Mild cognitive impairment; Cognition; Multivariate

Introduction

The thesis of this paper is that it is common for healthy adults and older adults to show "significant changes" in cognitive functioning when tested a second time. This issue has rarely been studied and it has major implications for clinical practice and research. Repeated cognitive testing is used for a variety of reasons, such as to evaluate for treatment effects, monitor recovery from a traumatic brain injury, or determine if a person has experienced a decline in functioning associated with a neurodegenerative disease, such as Alzheimer's disease. However, the rate at which people obtain cognitive test scores that are reliably improved or declined on retesting, in the absence of a change in clinical condition, is largely unknown and likely much greater than would be predicted from univariate estimates involving the standard normal distribution (i.e., the bell

curve). Simply put, clinicians and researchers do not have adequate data to determine how common it is to show improvement or decline on cognitive test scores in healthy adults and older adults who undergo retesting.

There are well established methods for interpreting change on cognitive testing, such as the reliable change methodology (Barr & McCrea, 2001; Chelune, Naugle, Luders, Sedlak, & Awad, 1993; Heaton et al., 2001; Hinton-Bayre, Geffen, Geffen, McFarland, & Friis, 1999; Iverson, 1998, 1999, 2001; Iverson, Lovell, & Collins, 2003; Iverson, 2012; Temkin, Heaton, Grant, & Dikmen, 1999) and standardized linear regression (McSweeney, Naugle, Chelune, & Luders, 1993; Salinsky, Storzbach, Dodrill, & Binder, 2001; Sawrie, Chelune, Naugle, & Luders, 1996; Sawrie, Marson, Boothe, & Harrell, 1999; Temkin et al., 1999). These methodologies are important and useful because they allow clinicians and researchers to estimate the probable range of measurement error surrounding test-retest difference scores, and then draw inferences about whether an individual patient or subject has experienced a statistically reliable improvement or worsening in performance. A limitation with using these methods in clinical practice is that they are calculated on the basis of a single test-retest score, and inherently assume that only one test has been administered (i.e., each change score is considered in isolation as if it was a single statistical analysis). In clinical practice and research, however, multiple tests are administered and interpreted. Therefore, these methods might accurately estimate the likelihood of change if a single test is administered and interpreted, but they will under-estimate the likelihood of change if multiple tests are given. For example, Iverson and colleagues illustrated that it is uncommon to have a statistically reliable worsening on computerized cognitive testing when only a single score is considered, but when several scores are considered simultaneously it is fairly common for healthy people to obtain at least one reliably lower score on a brief computerized battery (Iverson et al., 2003). Woods and colleagues also illustrated that multiple change scores need to be considered on retest (Woods et al., 2006). When considering retest scores for all 14 neuropsychological variables in healthy adults and adults with medically stable human immunodeficiency virus (HIV), Woods et al. reported that it was uncommon to have 7 or more scores reliably improve or reliably decline on retest in both controls and the clinical group.

The purpose of this study is to illustrate the prevalence of statistically reliable change scores on memory testing in healthy adults and older adults. We hypothesized that having at least one large improvement or decline on memory testing would be common when applying the reliable change methodology to several test scores simultaneously.

Methods

Participants

The participants for this study were from the test-retest samples in the standardization process for their respective published memory batteries. These batteries were the (a) Neuropsychological Assessment Battery, Memory Domain (NAB; Stern & White, 2003) and (b) Wechsler Memory Scale, Fourth Edition (WMS-IV; Wechsler, 2009).

Measures

Test-retest data were considered for all 10 NAB Memory subtest scores (List Learning A Immediate Recall, List Learning B Immediate Recall, List Learning A Short Delayed Recall, List Learning A Long Delayed Recall, Shape Learning Immediate Recognition, Shape Learning Delayed Recognition, Story Learning Phrase Unit Immediate Recall, Story Learning Phrase Unit Delayed Recall, Daily Living Memory Immediate Recall, and Daily Living Memory Delayed Recall) in the entire retest sample. The NAB retest sample also administered the Reynolds Intellectual Screening Test (RIST; Reynolds & Kamphaus, 2003).

For the WMS-IV, test-retest data were considered separately for the adult battery (16–69 years) and the older adult battery (65–90 years) due to differences in the subtests administered. For ages 16–69, all 5 WMS-IV index scores (Auditory Memory Index, Visual Memory Index, Immediate Memory Index, Delayed Memory Index, and Visual Working Memory Index) and 10 WMS-IV primary subtest scores (Logical Memory I, Logical Memory II, Verbal Paired Associates I, Verbal Paired Associates II, Designs I, Designs II, Visual Reproduction I, Visual Reproduction II, Spatial Addition, and Symbol Span) were considered. For ages 65–90, four WMS-IV index scores (Auditory Memory Index, Visual Memory Index, Immediate Memory Index, and Delayed Memory Index; the older adult battery does not include a Visual Working Memory Index) and seven primary WMS-IV subtest scores (Logical Memory I, Logical Memory II, Verbal Paired Associates I, Verbal Paired Associates II, Visual Reproduction I, Visual Reproduction II, and Symbol Span) were considered. The WMS-IV retest sample was also administered the Wechsler Adult Intelligence Scale, Fourth Edition (Wechsler, 2008).

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 [SDs] at Time 1 and Time 2, and test-retest correlations). For this study, we used RCI with a 90% confidence interval (CI) and included practice effects. Practice effects are defined as the difference in mean group scores between Time 2 and Time 1, which is subsequently added to the RCI. The steps for calculating reliable change scores are provided below and Table 1 shows the values used in this study.

(1) $SEM_1 = SD\sqrt{1 - r_{12}}$ SD from Time 1 multiplied by the square root of 1 minus the test-retest coefficient.
(2) $SEM_2 = SD\sqrt{1 - r_{12}}$ SD from Time 2 multiplied by the square root of 1 minus the test-retest coefficient.
(3) $RCI = \sqrt{SEM_1^2 + SEM_2^2}$ Square root of the sum of the squared SEMs for each testing occasion.
(4) RCI 90%CI = RCI × 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 each battery, including the prevalence of 0, 1, 2, and so forth. reliable change scores when all change scores are 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 (a) participant education ($\leq$12 years or >12 years), (b) level of overall intellectual abilities ($\leq$50th percentile or >50th percentile), (c) sex (male or female), and (d) ethnicity (Caucasian or not Caucasian [i.e., African American, Asian, Hispanic, and Other]) to further evaluate the key principles of multivariate base rates that have been previously reported (Brooks, Strauss, Sherman, Iverson, & Slick, 2009b; Brooks & Iverson, 2012; Brooks, Iverson, & Holdnack, 2013; Iverson & Brooks, 2011). Further analyses were carried out to determine whether those healthy adults and older adults who would be deemed to have memory impairment based solely on test performance (e.g., one or more memory scores <1.5 SDs below the mean; did not include working memory scores) were more or less likely to have reliably changes scores on retest. Analyses to determine whether frequencies differed across any groups were done using chi-square analyses with odds ratios (ORs) being presented when results are significant. Due to the large number of analyses and potential for Type I errors, alpha was set at $p < .01$.

Results

Descriptions of the test-retest samples are presented in Table 2. Mean retest intervals for the adult memory batteries ranged from about 3 weeks on the WMS-IV to approximately 6 months on the NAB Memory Module. Samples are roughly evenly split between sexes and have a higher proportion of Caucasians. The frequencies of reliable change scores, when considering each score in isolation (univariate analyses), are presented at the bottom of Table 2. The mean percentages of healthy people who obtain a reliably different score on retest are approximately at the expected 10% range for both indexes and subtests, and are generally symmetrical for declined (approximately 5%) and improved (approximately 5%).

The multivariate frequencies of reliable change subtest scores on the NAB Memory Module are presented in Table 3. In this retest sample, 57.9% had one or more reliably different subtest scores in any direction. In the NAB retest sample, 31.6% had one or more reliably declined subtest scores and 33.7% had one or more reliably improved subtest scores [$\chi^2(1) = 0.77$, $p = .38$]. Those with higher intellectual abilities (RIST > 100) were more likely to have a reliable decline on retest compared to those with lesser intellectual abilities (RIST $\leq$ 100) [$\chi^2(1) = 13.68$, $p < .001$, OR = 7.0 (95%CI = 2.1–24.2)]. Those with lesser intellectual abilities (RIST $\leq$ 100) were not more likely to have one or more reliably improved NAB Memory subtest scores compared to those with higher intellectual abilities [RIST > 100; $\chi^2(1) = 1.48$, $p = .22$]. When stratifying by education level (i.e., 12 or fewer years vs. more than 12 years), there were no significant differences in the prevalence rates of reliably declined [$\chi^2(1) = 0.01$, $p = .96$] or reliably improved [$\chi^2(1) = 0.08$, $p = .78$] subtest scores. There were not differences in the NAB subtest reliable change base rates between males and females [declined: $\chi^2(1) = 5.23$, $p = .02$; improved: $\chi^2(1) = 1.04$, $p = .31$] or between Caucasian and non-Caucasian [declined: $\chi^2(1) = 0.04$, $p = .55$; improved: $\chi^2(1) = 0.14$, $p = .71$].

Multivariate base rates of reliable change scores for the WMS-IV are presented in Table 4. The occurrence of one or more reliably changed index scores (any direction) was found in 26.1% of the WMS-IV 16–69-year-old retest sample. In this

Table 1. Reliable change scores used for the NAB and WMS-IV

Subtest and index names	SD$_1$	SD$_2$	r$_{12}$	SEM$_1$	SEM$_2$	SE$_{diff}$	Mean PE	90%CI Declined	90%CI Improved
NAB									
Memory Index	13.8	14.2	0.67	7.9	8.2	11.4	5.6	−14	+25
List Learning List A Immediate Recall	9.0	9.4	0.34	7.3	7.6	10.6	1.9	−16	+20
List Learning List B Immediate Recall	10.1	10.1	0.40	7.8	7.8	11.1	−1.5	−20	+17
List Learning List A Short Delayed Recall	9.3	9.8	0.41	7.1	7.5	10.4	1.3	−16	+19
List Learning List A Long Delayed Recall	9.5	8.7	0.39	7.4	6.8	10.1	2.8	−14	+20
Shape Learning Immediate Recognition	10.2	11.4	0.43	7.7	8.6	11.5	3.7	−16	+23
Shape Learning Delayed Recognition	10.1	11.2	0.40	7.8	8.7	11.7	2.0	−18	+22
Story Learning Immediate Recall	9.8	8.5	0.58	6.4	5.5	8.4	2.3	−12	+17
Story Learning Delayed Recall	9.0	8.9	0.57	5.9	5.8	8.3	2.3	−12	+16
Daily Living Memory Immediate Recall	10.0	9.2	0.49	7.1	6.6	9.7	4.8	−12	+21
Daily Living Memory Delayed Recall	10.8	10	0.53	7.4	6.9	10.1	1.4	−16	+18
WMS-IV, 16–69 year battery									
Logical Memory I	2.9	2.6	0.72	1.5	1.4	2.1	1.9	−2	+6
Logical Memory II	2.8	2.9	0.67	1.6	1.7	2.3	2.3	−2	+7
Verbal Paired Associates I	3.1	3.4	0.76	1.5	1.7	2.3	2.3	−2	+6
Verbal Paired Associates II	3.0	2.7	0.76	1.5	1.3	2.0	1.0	−3	+5
Verbal Paired Associates II Word Recall	2.9	3.2	0.74	1.5	1.6	2.2	0.9	−3	+5
Designs I	2.9	3.4	0.73	1.5	1.8	2.3	1.1	−3	+5
Designs II	2.7	3.2	0.72	1.4	1.7	2.2	1.7	−2	+6
Visual Reproduction I	2.8	2.8	0.62	1.7	1.7	2.4	1.9	−3	+6
Visual Reproduction II	2.8	3.0	0.59	1.8	1.9	2.6	2.8	−2	+8
Spatial Addition	2.8	3.0	0.74	1.4	1.5	2.1	0.8	−3	+5
Symbol Span	3.0	3.1	0.72	1.6	1.6	2.3	0.6	−4	+5
Auditory Memory Index	14.1	14.4	0.81	6.1	6.3	8.8	11.5	−3	+26
Visual Memory Index	14.8	16.6	0.8	6.6	7.4	9.9	12.1	−5	+29
Visual Working Memory Index	14.4	15.6	0.82	6.1	6.6	9.0	4.3	−11	+20
Immediate Memory Index	14.9	15.6	0.81	6.5	6.8	9.4	12.4	−3	+28
Delayed Memory Index	13.9	15	0.79	6.4	6.9	9.4	13.7	−2	+30
WMS-IV, 65–90 year battery									
Logical Memory I	2.9	3.3	0.77	1.4	1.6	2.1	2.0	−2	+6
Logical Memory II	2.7	2.8	0.71	1.5	1.5	2.1	2.1	−2	+6
Verbal Paired Associates I	2.8	2.9	0.76	1.4	1.4	2.0	1.7	−2	+5
Verbal Paired Associates II	2.7	2.7	0.77	1.3	1.3	1.8	1.1	−2	+5
Verbal Paired Associates II Word Recall	2.8	2.7	0.72	1.5	1.4	2.1	1.2	−3	+5
Visual Reproduction I	3.1	3.1	0.79	1.4	1.4	2.0	1.8	−2	+6
Visual Reproduction II	2.8	3.1	0.64	1.7	1.9	2.5	1.8	−3	+6
Symbol Span	2.8	3.0	0.69	1.6	1.7	2.3	0.6	−4	+5
Auditory Memory Index	12.9	14.5	0.82	5.5	6.2	8.2	10.6	−3	+25
Visual Memory Index	14.6	17.0	0.79	6.7	7.8	10.3	11.0	−6	+28
Immediate Memory Index	13.9	14.2	0.84	5.6	5.7	7.9	12.4	−1	+26
Delayed Memory Index	13.1	14.4	0.80	5.9	6.4	8.7	11.0	−4	+26

Notes: SD = standard deviation; r$_{12}$ = retest correlation; SEM = standard error of measurement; PE = practice effect (mean Time 2−mean Time 1); NAB = Neuropsychological Assessment Battery (Stern and White, 2003); WMS-IV = Wechsler Memory Scale, Fourth Edition (Wechsler, 2009).

sample, 11.9% had one or more reliably declined index scores and 15.7% had one or more reliably improved index scores [$\chi^2(1) = 0.53$, $p = 0.47$]. There were no significant differences in the base rates of reliably declined [$\chi^2(1) = 4.3$, $p = .04$] or reliably improved [$\chi^2(1) = 0.03$, $p = .86$] index scores when considering intellectual abilities that are either WAIS-IV FSIQ ≤ 100 or WAIS-IV FSIQ > 100. When stratifying by education level (i.e., 12 or fewer years vs. more than 12 years), there were no significant differences in the prevalence rates of reliably declined [15.4% and 8.7%, respectively; $\chi^2(1) = 1.42$, $p = .23$] or reliably improved [10.8% and 20.3%, respectively; $\chi^2(1) = 2.30$, $p = .13$] index scores. There were not differences in the

Table 2. Test-retest sample characteristics for the adult memory batteries

Demographics	NAB memory		WMS-IV			
			Adult		Older adult	
N	95		173		71	
Mean age, years (SD)	50.5 (17.6)		40.7 (17.0)		75.6 (7.4)	
Mean retest interval, weeks (range)	27.6 (22.3–37.7)		3.2 (2.0–12.0)		3.2 (2.0–12.0)	
Sex (% female)	66		51		59	
Ethnicity (% Caucasian)	73		69		85	
Univariate frequencies of reliable change scores	Decline	Improve	Decline	Improve	Decline	Improve
Indexes						
Mean percent of sample	2.2	3.3	3.9	5.2	3.6	6.1
SD	—[a]	—	0.8	0.6	1.8	2.5
Range	—	—	3.0–4.7	4.5–6.0	1.4–5.6	2.9–8.7
Subtests						
Mean percent of sample	6.9	7.4	4.5	5.4	3.0	7.1
SD	3.7	3.0	2.0	2.2	1.3	3.5
Range	0–11.6	3.2–11.6	2.3–7.9	2.6–8.8	1.4–4.3	2.8–12.7

Notes: Univariate frequencies of reliable change scores represent the mean percentage of the sample identified as reliably declined or reliably improved if considering each change score in isolation. For example, 3.0–4.7% of adults showed a reliable decline on each of the five WMS-IV indexes, with and average decline of 3.9% (and expected base rate of 5%, based on the 90%CI for reliable change).
[a]NAB Memory yields only one index score, so multivariate analyses cannot be computed. SD = standard deviation.

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

Number of subtest scores	Total sample	RIST < 100	RIST ≥ 100	Education ≤12 years	Education >12 years
Cumulative percentages with number of reliably declined NAB memory subtest scores					
Zero	68.4	82.4	59.3	68.0	68.6
1 or more	31.6	11.8	40.7	32.0	31.4
2 or more	10.5	2.9	13.6	20.0	7.1
3 or more	2.1	2.9	1.7	4.0	1.4
4 or more	2.1	2.9	1.7	4.0	1.4
5 or more	2.1	2.9	1.7	4.0	1.4
6 or more	2.1	2.9	1.7	4.0	1.4
Cumulative percentages with number of reliably improved NAB memory subtest scores					
Zero	66.3	58.8	71.2	64.0	67.1
1 or more	33.7	41.2	28.8	36.0	32.9
2 or more	8.4	17.6	3.4	8.0	8.6
3 or more	3.2	5.9	1.7	—	4.3
4 or more	1.1	2.9	—	—	1.4

Notes: RIST = Reynolds Intellectual Screening Test (Reynolds & Kamphaus, 2003). All 10 NAB Memory subtest scores were considered (List Learning A Immediate Recall, List Learning B Immediate Recall, List Learning A Short Delayed Recall, List Learning A Long Delayed Recall, Shape Learning Immediate Recognition, Shape Learning Delayed Recognition, Story Learning Phrase Unit Immediate Recall, Story Learning Phrase Unit Delayed Recall, Daily Living Memory Immediate Recall, and Daily Living Memory Delayed Recall).

WMS-IV Index base rates between males and females [declined: $\chi^2(1) = 1.99$, $p = .16$; improved: $\chi^2(1) = 0.21$, $p = .64$] or between Caucasian and non-Caucasian [declined: $\chi^2(1) = 0.84$, $p = .36$; improved: $\chi^2(1) = 0.01$, $p = .98$].

    For the WMS-IV subtests, 62.5% of this sample had one or more reliably different scores in any direction. One or more reliably declined subtest scores was found in 33.6% and one or more reliably improved subtest scores was found in 39.1% of

Table 4. Cumulative percentages of the WMS-IV adult sample with reliably different scores on retest when considering all change scores simultaneously

	Total Sample	WAIS-IV FSIQ < 100	WAIS-IV FSIQ ≥ 100	Education ≤12 years	Education >12 years
Cumulative percentages with number of reliably declined WMS-IV scores					
Number of index scores					
Zero	88.1	81.8	93.8	84.6	91.3
1 or more	11.9	18.2	6.3	15.4	8.7
2 or more	5.2	6.1	4.7	6.2	4.3
3 or more	0.7	—	1.6	—	1.4
4 or more	—	—	—	—	—
Number of subtest scores					
Zero	66.4	57.4	74.6	60.0	72.1
1 or more	33.6	42.6	25.4	40.0	27.9
2 or more	9.4	16.4	3.2	13.3	5.9
3 or more	1.6	—	3.2	1.7	1.5
Cumulative percentages with number of reliably improved WMS-IV scores					
Number of index scores					
Zero	84.3	84.8	85.9	89.2	79.7
1 or more	15.7	15.2	14.1	10.8	20.3
2 or more	6.0	4.5	6.3	3.1	8.7
3 or more	2.2	3.0	—	—	4.3
4 or more	0.7	—	—	—	1.4
Number of subtest scores					
Zero	60.9	63.9	58.7	63.3	58.8
1 or more	39.1	36.1	41.3	36.7	41.2
2 or more	9.4	8.2	9.5	6.7	11.8
3 or more	3.1	3.3	3.2	1.7	4.4

Notes: For ages 16–69, all five WMS-IV index scores (Auditory Memory Index, Visual Memory Index, Immediate Memory Index, Delayed Memory Index, and Visual Working Memory Index) and 10 WMS-IV primary subtest scores (Logical Memory I, Logical Memory II, Verbal Paired Associates I, Verbal Paired Associates II, Designs I, Designs II, Visual Reproduction I, Visual Reproduction II, Spatial Addition, and Symbol Span) were considered.

the sample [$\chi^2(1) = 0.83$, $p = .36$]. There were no significant differences in the base rates of reliably declined [$\chi^2(1) = 4.11$, $p = 0.04$] or reliably improved [$\chi^2(1) = 0.35$, $p = .55$] subtest scores when considering intellectual abilities that are either WAIS-IV FSIQ ≤ 100 or WAIS-IV FSIQ > 100. There were not differences in the WMS-IV subtest base rates between males and females [declined: $\chi^2(1) = 2.83$, $p = .09$; improved: $\chi^2(1) = 0.62$, $p = .43$] or between Caucasian and non-Caucasian [declined: $\chi^2(1) = 0.57$, $p = .45$; improved: $\chi^2(1) = 0.03$, $p = .88$].

In the 65–90 year olds who were given the older adult battery, 25.0% of the WMS-IV sample had one or more reliably changed index scores on retest (Table 5). In this older adult sample, 7.4% had one or more reliably declined index scores and 19.1% had one or more reliably improved index scores [$\chi^2(1) = 4.10$, $p = .04$]. There were no significant differences in the base rates of reliably declined WMS-IV index scores in those older adults with lesser versus higher intellectual abilities [$\chi^2(1) = 0.10$, $p = .75$]. However, those older adults with higher intellectual abilities (WAIS-IV FSIQ > 100) were more likely to have one or more reliably improved WMS-IV index scores compared to those with lesser intellectual abilities [WAIS-IV FSIQ ≤ 100; $\chi^2(1) = 9.05$, $p = .003$, OR not computed due to zero count in one cell]. When stratifying by education level (i.e., 12 or fewer years vs. more than 12 years), there were no significant differences in the prevalence rates of reliably declined index scores [10.3% and 3.4%, respectively; $\chi^2(1) = 1.13$, $p = .29$] or reliably improved index scores on retest [31.0% and 13.2%, respectively; $\chi^2(1) = 4.64$, $p = .03$]. There were not differences in the WMS-IV Index base rates for older adults when comparing males and females [declined: $\chi^2(1) = 0.78$, $p = .38$; improved: $\chi^2(1) = 1.37$, $p = .24$] or when comparing Caucasian and non-Caucasian [declined: $\chi^2(1) = 0.21$, $p = .64$; improved: $\chi^2(1) = 0.53$, $p = .47$].

In the 65–90-year-old sample, 52.2% had one or more reliably different WMS-IV subtest scores in any direction. One or more reliably declined subtest scores was found in 13.4% of this sample, but older adults were significantly more likely to have one or more reliably improved subtest scores [43.3% of the sample; $\chi^2(1) = 14.69$, $p < .001$, OR = 4.9 (95%CI = 2.0–12.7)]. When stratifying by education level (i.e., 12 or fewer years vs. more than 12 years), there were no significant differences in the prevalence rates of reliably declined [40.0% and 27.9%, respectively; $\chi^2(1) = 2.08$, $p = .15$] or reliably improved [36.7% and 41.2%, respectively; $\chi^2(1) = 0.27$, $p = .60$] subtest scores. There were not significant differences in the base rates of reliably

Table 5. Cumulative percentages of the WMS-IV older adult sample with reliably different scores on retest when considering all change scores simultaneously

	Total Sample	WAIS-IV FSIQ < 100	WAIS-IV FSIQ ≥ 100	Education ≤12 years	Education >12 years
Cumulative percentages with number of reliably *declined* WMS-IV scores					
Number of index scores					
Zero	92.6	95.2	93.1	89.7	96.6
1 or more	7.4	4.8	6.9	10.3	3.4
2 or more	1.5	—	—	2.6	—
3 or more	1.5	—	—	—	—
4	1.5	—	—	—	—
Number of subtest scores					
Zero	86.6	85.7	89.7	89.7	69.0
1 or more	13.4	14.3	10.3	10.3	31.0
2 or more	4.5	4.8	6.9	2.6	10.3
3 or more	1.5	4.8	—	—	—
Cumulative percentages with number of reliably *improved* WMS-IV scores					
Number of index scores					
Zero	80.9	100	65.5	86.8	86.2
1 or more	19.1	0	34.5	13.2	13.8
2 or more	5.9	—	13.8	5.3	3.4
3 or more	—	—	—	2.6	—
4	—	—	—	—	—
Number of subtest scores					
Zero	56.7	76.2	37.9	55.3	58.6
1 or more	43.3	23.8	62.1	44.7	41.4
2 or more	6.0	—	6.9	2.6	10.3
3 or more	1.5	—	—	—	3.4

Notes: For ages 65–90, four WMS-IV index scores (Auditory Memory Index, Visual Memory Index, Immediate Memory Index, and Delayed Memory Index) and seven primary WMS-IV subtest scores (Logical Memory I, Logical Memory II, Verbal Paired Associates I, Verbal Paired Associates II, Visual Reproduction I, Visual Reproduction II, and Symbol Span) were considered.

declined WMS-IV subtest scores for older adults based on intellectual abilities [$\chi^2(1) = 0.18$, $p = .67$]. Older adults with higher intellectual abilities were statistically more likely to have one or more reliably improved WMS-IV subtest scores compared to older adults with lesser intellectual abilities [$\chi^2(1) = 7.18$, $p = .007$, OR = 5.2 (95%CI = 1.3–22.6)]. When stratifying by education level (i.e., 12 or fewer years vs. more than 12 years), there were no significant differences in the prevalence rates of reliably declined [13.2% and 13.8%, respectively; $\chi^2(1) = 0.01$, $p = .94$] or reliably improved [44.7% and 41.4%, respectively; $\chi^2(1) = 0.08$, $p = .78$] subtest scores. There were not differences in the WMS-IV subtest base rates for older adults when comparing males and females [declined: $\chi^2(1) = 0.09$, $p = .77$; improved: $\chi^2(1) = 0.34$, $p = .56$] or when comparing Caucasian and non-Caucasian [declined: $\chi^2(1) = 0.01$, $p = .99$; improved: $\chi^2(1) = 0.8$, $p = .78$].

The rates of reliable change scores in those deemed to have low memory scores (i.e., "memory impairment") are presented in Table 6. For 16–69 year olds, memory impairment is found in 27.3% at Time 1 and is found in 13.3% at Time 2. For 65–90 year olds, memory impairment is found in 14.9% at Time 1 and in 9.0% at Time 2. There were no significant differences in the prevalence rates of change scores depending on how memory impairment was defined (e.g., considering an isolated low score or considering multiple low scores; all ps > .05), so results are presented using one or more scores <1.5SDs in order to maximize group sizes. In adults on the WMS-IV, those with "memory impairment" were not more likely than those without memory impairment to have one or more scores reliably decline on retest [42.9% vs. 30.1%; $\chi^2(1) = 1.85$, $p = .17$] or reliably improve on retest [54.3% vs. 33.3%; $\chi^2(1) = 4.69$, $p = .03$]. In older adults on the WMS-IV, those with "memory impairment" were not more likely than those without memory impairment to have one or more scores reliably decline on retest [20.0% vs. 12.3%; $\chi^2(1) = 0.44$, $p = .51$] or reliably improve on retest [30.0% vs. 45.6%; $\chi^2(1) = 0.85$, $p = .36$]. In both adults and older adults, those with "memory impairment" had similar rates of showing a reliable decline on retest as they did a reliable improvement on retest (all ps > .05). For older adults, those who did not have memory impairment were more likely to show improvement than to show decline when retested [$\chi^2(1) = 15.40$, $p < .001$, OR = 6.0 (95%CI = 2.1–17.4)].

Table 6. Base rates of reliable change subtest scores for those who do or do not meet criteria for "memory impairment"

Number of subtest scores that reliably change on retest	Meets criteria for "Memory Impairment"		Does not meet criteria for "Memory Impairment"	
	Reliably declined	Reliably improved	Reliably declined	Reliably improved
Adults (16–69 years)				
Zero	57.1	45.7	69.9	66.7
1 or more	42.9	54.3	30.1	33.3
2 or more	17.1	14.3	6.5	7.5
3 or more	2.9	5.7	1.1	2.2
Older adults (65–90 years)				
Zero	80.0	70.0	87.7	54.4
1 or more	20.0	30.0	12.3	45.6
2 or more	—	—	5.3	7.0
3 or more	—	—	1.8	1.8

Note: SD = standard deviation. "Memory impairment" was defined as having one or more subtest scores that were more than 1.5 SDs below the mean. Note, however, that these are considered normal control subjects with low memory score, not a clinical sample. Analyses are based on the WMS-IV standardization sample. Standardization data from the *Wechsler Memory Scale, Fourth Edition (WMS-IV)*. Copyright © 2009 NCS Pearson, Inc. Used with permission. All rights reserved.

Discussion

This study illustrated persuasively that it is common for people to have at least one large change in performance on memory testing when several memory tests are administered. The changes in performance were not unidirectional and reflective of practice effects; they were bidirectional. This finding was present in healthy adults and older adults, and occurred for normative scores that were both age adjusted and demographically adjusted. These results are consistent with prior studies illustrating that it is common to have one or more reliable change scores on retesting (Iverson et al., 2003; Woods et al., 2006), and have important implications for clinical practice and research when trying to determine if a patient or research participant has shown improvement or worsening in memory functioning over time.

Multivariate base rates of reliable change scores far exceed the expectations from interpretation of single reliable change scores. When considering a single reliable change score in isolation, a 90%CI around the change score (adjusted for practice effects) should result in approximately 10% of healthy people showing a reliable difference when retested (i.e., ~5% showing a reliable decline and ~5% showing a reliable improvement; as shown in Table 2). Across the memory batteries in this study, the prevalence rates for having one or more reliable change scores (in either direction) was 25–26% for index scores and ranged from 52% to 63% for subtest scores. Having one or more reliable declines in performance was found in 7–12% for index scores and 13–34% for subtest scores. Having one or more reliable improvements was found in 16–19% for index scores and 34–43% for subtest scores across samples and batteries. In general, prevalence rates for reliable change scores on both indexes and subtests become uncommon with two or more reliably declined or two or more reliably improved scores, and in some scenarios it is only uncommon to have three or more reliably changed retest scores. Clearly, considering multiple reliable change scores can easily result in misinterpretation of retest performance if a clinician is expecting univariate prevalence rates.

These results are related to and align with other multivariate interpretive considerations for cognitive testing. There is now a large and mature body of evidence illustrating that when a battery of cognitive test scores are interpreted, it is common for healthy adults to obtain one or more unusually low scores (Brooks & Iverson, 2010; Brooks, Iverson, & White, 2009a; Brooks, Holdnack, & Iverson, 2011; Brooks et al., 2013; Iverson, Brooks, White, & Stern, 2008). Similarly, within a single domain of cognitive functioning, such as working memory (Brooks et al., 2013; Iverson, Brooks, & Holdnack, 2012), episodic memory (Brooks, Iverson, Holdnack, & Feldman, 2008; Brooks, Iverson, & White, 2007; Brooks et al., 2009a), executive functioning (Brooks, Iverson, Lanting, Horton, & Reynolds, 2012), and processing speed (Brooks et al., 2013; Iverson et al., 2012), a substantial minority of healthy adults and older adults will obtain one low score. The prevalence of low scores when simultaneously interpreting multiple scores is attributable to several factors, including: the number of scores being interpreted (e.g., there is a direct relation between the number of scores interpreted and the number of expected low scores or reliable change scores obtained); the cutoff score being used for interpretation (e.g., more lenient cutoff scores will produce more low scores or more reliable change scores); and the strength of the intercorrelations between the scores being considered (e.g., lower intercorrelations between tests will produce higher rates of low scores or higher rates of reliable change scores).

These multivariate interpretation studies also illustrate that the rate of obtaining low scores varies by race, level of education, and level of intellectual ability (for discussions, see Brooks et al., 2009b; Brooks & Iverson, 2012; Brooks et al., 2013; Iverson & Brooks, 2011), which was partially replicated when considering these reliable change scores (e.g., those with higher intellectual abilities were more likely to show reliable change on retest, but education, sex, and ethnicity were not factors in the prevalence of change scores for these samples). It is also expected that when clinicians interpret multiple index-level discrepancy scores (for example, discrepancies between intelligence and memory or verbal intelligence vs. processing speed), the base rate of having one unusual discrepancy is considerably greater than is reported in test manuals and computer scoring programs because those base rates consider the prevalence of each discrepancy in isolation (not considering multiple discrepancies simultaneously, as is usually done in clinical practice) (Crawford, Allum, & Kinion, 2008; Crawford, Anderson, Rankin, & Macdonald, 2010). These studies on the multivariate base rates of low scores, reliable change scores, and significant discrepancy scores have considerable implications for how we identify cognitive impairment, how we track potential cognitive decline, and how accurately neuropsychological testing can contribute to the diagnosis of disease (Ganguli et al., 2011; Jak et al., 2009; Jak et al., 2009; Litvan et al., 2012; Schinka et al., 2010). Additional research is needed to drill deeper into these findings and better understand factors, other than a change in actual functioning, that are related to large test-retest changes on cognitive test scores. Prior research has supported the use of Monte Carlo simulation for multivariate base rates of low scores (Brooks & Iverson, 2010), so it is possible that this method could prove to be powerful for better understanding these psychometric factors when interpreting large amounts of change scores.

There are some limitations to this study that should be considered. The samples used in this study were fairly small, and we could not stratify them by multiple levels of education, intellectual ability, or ethnicity. Therefore, we could not adequately explore the relationship between these factors and the likelihood of having large change scores. 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 readily calculated by clinicians and researchers for any score using the test-retest information presented in technical manuals. Other methods, such as regression, require access to the raw standardization data. Nonetheless, regression modeling could be used to better explore additional factors relating to change scores. 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 subsequently used to evaluate the multivariate base rates of these reliable change scores. Although this is the same process as has been used in many past studies involving base rates of low test scores (i.e., standardization sample used to derive the test scores, then base rates considered in the same data), having second independent samples to validate the results may be helpful. And finally, it is important to consider that the test-retest intervals of these standardized measures (mean number of weeks ranging from 3 to 28) may not directly reflect retest intervals for clinical assessments. This does not mean that the information provided by this study is dubious, but it does suggest that generalizability could be limited and further research is needed on the multivariate base rates of reliable change scores with longer retest intervals.

Determining whether there has been improvement or worsening in cognitive functioning is a complicated process that involves a series of steps and an understanding of several psychometric concepts. Variability in scores over time might relate to measurement error, situational factors, and practice effects—and changes in test performance may or may not reflect true changes in cognitive functioning. The clinician or researcher must consider carefully the potential influences of (a) initial level of performance, (b) practice effects, (c) regression to the mean, (d) measurement error, and (e) motivational factors (see Iverson, 2012 for a discussion of these issues). One of the most important factors to consider is level of performance during initial testing. The effects of practice and regression vary in relation to initial test scores. Low scores are likely to be higher on retesting due to situational factors, practice, and regression to the mean. High scores are less likely to improve due to practice and, in contrast, they are more likely to worsen due to regression and situational factors. The present study illustrates that a large improvement or decline in memory test performance is much more likely to occur when several tests are administered, as is usually the case in clinical practice.

Funding

Conflict of Interest

Brian Brooks receives royalties for tests published by Psychological Assessment Resources, Inc. (i.e., Child and Adolescent Memory Profile [ChAMP], Sherman and Brooks, 2015a; Memory Validity Profile [MVP], Sherman and Brooks, 2015b), which is also the publisher of the Neuropsychological Assessment Battery (Stern and White, 2003). 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 WAIS-IV (Wechsler, 2008) and WMS-IV (Wechsler, 2009). He receives royalties from Elsevier Science for the WAIS-IV/WMS-IV/ACS Advanced Interpretation book. 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 (not in the past 5 years). He receives royalties from books in neuropsychology and one neuropsychological test (Psychological Assessment Resources, Inc.).

Acknowledgements

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* (*6*), 693–702.

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., & Feldman, H. H. (2008). Potential for misclassification of mild cognitive impairment: A study of memory scores on the Wechsler Memory Scale-III in healthy older adults. *Journal of the International Neuropsychological Society*, *14* (*3*), 463–478.

Brooks, B. L., Iverson, G. L., Lanting, S. C., Horton, A. M., & Reynolds, C. R. (2012). Improving test interpretation for detecting executive dysfunction in adults and older adults: prevalence of low scores on the test of verbal conceptualization and fluency. *Applied Neuropsychology*, *19* (*1*), 61–70.

Brooks, B. L., Iverson, G. L., & White, T. (2007). Substantial risk of "Accidental MCI" in healthy older adults: Base rates of low memory scores in neuropsychological assessment. *Journal of the International Neuropsychological Society*, *13* (*3*), 490–500.

Brooks, B. L., Iverson, G. L., & White, T. (2009a). Advanced interpretation of the Neuropsychological Assessment Battery (NAB) with older adults: Base rate analyses, discrepancy scores, and interpreting change. *Archives of Clinical Neuropsychology*, *24* (*7*), 647–657.

Brooks, B. L., Strauss, E., Sherman, E. M. S., Iverson, G. L., & Slick, D. J. (2009b). Developments in neuropsychological assessment: Refining psychometric and clinical interpretive methods. *Canadian Psychology*, *50* (*3*), 196–209.

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., & Iverson, G. L. (2012). Improving accuracy when identifying cognitive impairment in pediatric neuropsychological assessments. In E. M. S., Sherman, & B. L., Brooks (Eds.), *Pediatric forensic neuropsychology* (pp. 66–88). New York: Oxford University Press.

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

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.

Crawford, J. R., Allum, S., & Kinion, J. E. (2008). An index-based short form of the WAIS-III with accompanying analysis of reliability and abnormality of differences. *British Journal of Clinical Psychology*, *47* (*Pt 2*), 215–237.

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.

Ganguli, M., Snitz, B. E., Saxton, J. A., Chang, C. C., Lee, C. W., & Vander Bilt, J. (2011). Outcomes of mild cognitive impairment by definition: A population study. *Archives of Neurology*, *68* (*6*), 761–767.

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* (*1*), 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* (*10*), 661–666.

Iverson, G. L.1999Interpreting 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., Brooks, B. L., White, T., & Stern, R. A. (2008). Neuropsychological Assessment Battery (NAB): Introduction and advanced interpretation. In A. M., Horton Jr., & D., Wedding (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* (*4*), 460–467.

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

Iverson, G. L., & Brooks, B. L. (2011). Improving accuracy for identifying cognitive impairment. In M. R., Schoenberg, & J. G., Scott (Eds.), *The black book of neuropsychology: A syndrome-based approach* (pp. 923–950). New York: Springer.

Iverson, G. L., Brooks, B. L., & Holdnack, J. A. (2012). Evidence-based neuropsychological assessment following work-related injury. In S. S., Bush, & G. L., Iverson (Eds.), *Neuropsychological assessment of workplace injuries*. New York: Guilford 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* (*1*), 12–19.

Jak, A. J., Bangen, K. J., Wierenga, C. E., Delano-Wood, L., Corey-Bloom, J., & Bondi, M. W. (2009). Contributions of neuropsychology and neuroimaging to understanding clinical subtypes of mild cognitive impairment. *International Reviews in Neurobiology*, *84*, 81–103.

Jak, A. J., Bondi, M. W., Delano-Wood, L., Wierenga, C., Corey-Bloom, J., & Salmon, D. P. (2009). Quantification of five neuropsychological approaches to defining mild cognitive impairment. *American Journal of Geriatric Psychiatry*, *17* (*5*), 368–375.

Litvan, I., Goldman, J. G., Troster, A. I., Schmand, B. A., Weintraub, D., & Petersen, R. C. (2012). Diagnostic criteria for mild cognitive impairment in Parkinson's disease: Movement disorder society task force guidelines. *Movement Disorder*, *27* (*3*), 349–356.

McSweeney, A. J., Naugle, R. I., Chelune, G. J., & Luders, H. (1993). "T scores for change": An illustration of a regression approach to depicting change in clinical neuropsychology. *Clinical Neuropsychologist*, *7*, 300–312.

Reynolds, C. R., & Kamphaus, R. W. (2003). *Reynolds intellectual assessment scales*. Lutz, FL: Psychological Assessment Resources, Inc.

Salinsky, M. C., Storzbach, D., Dodrill, C. B., & Binder, L. M. (2001). Test-retest bias, reliability, and regression equations for neuropsychological measures repeated over a 12–16-week period. *Journal of the International Neuropsychological Society*, *7* (*5*), 597–605.

Sawrie, S. M., Chelune, G. J., Naugle, R. I., & Luders, H. O. (1996). Empirical methods for assessing meaningful neuropsychological change following epilepsy surgery. *Journal of the International Neuropsychological Society*, *2* (*6*), 556–564.

Sawrie, S. M., Marson, D. C., Boothe, A. L., & Harrell, L. E. (1999). A method for assessing clinically relevant individual cognitive change in older adult populations. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *54* (*2*), P116–124.

Schinka, J. A., Loewenstein, D. A., Raj, A., Schoenberg, M. R., Banko, J. L., & Potter, H. (2010). Defining mild cognitive impairment: Impact of varying decision criteria on neuropsychological diagnostic frequencies and correlates. *American Journal of Geriatric Psychiatry*, *18* (*8*), 684–691.

Sherman, E. M. S., & Brooks, B. L. (2012). *Pediatric Forensic Neuropsychology*. New York: Oxford University Press.

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

Sherman, E. M. S., & Brooks, B. L. (2015b). *Memory Validity Profile (MVP)*. Lutz, FL: Psychological Assessment Resources, Inc.

Stern, R. A., & White, T. (2003). *Neuropsychological assessment battery*. 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* (*4*), 357–369.

Wechsler, D. (2008). *Wechsler adult intelligence scale* (4th ed.). San Antonio, TX: Pearson.

Wechsler, D. (2009). *Wechsler memory scale* (4th ed.). San Antonio, TX: Pearson.

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* (*1*), 83–89.