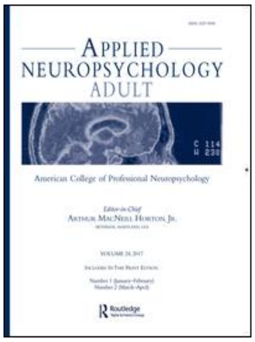

Applied Neuropsychology: Adult

ISSN: 2327-9095 (Print) 2327-9109 (Online) Journal homepage: https://www.tandfonline.com/loi/hapn21

Examining 3-month test-retest reliability and reliable change using the Cambridge Neuropsychological Test Automated Battery

Rune H. Karlsen, Justin E. Karr, Simen B. Saksvik, Astri J. Lundervold, Odin Hjemdal, Alexander Olsen, Grant L. Iverson & Toril Skandsen

APPLIED NEUROPSYCHOLOGY: ADULT
https://doi.org/10.1080/23279095.2020.1722126

Routledge
Taylor & Francis Group

OPEN ACCESS

Examining 3-month test-retest reliability and reliable change using the Cambridge Neuropsychological Test Automated Battery

Rune H. Karlsen[a], Justin E. Karr[b,c,d], Simen B. Saksvik[e,f], Astri J. Lundervold[g], Odin Hjemdal[e], Alexander Olsen[e,f], Grant L. Iverson[b,c,d], and Toril Skandsen[a,f]

[a]Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, Trondheim, Norway; [b]Departments of Psychiatry and Physical Medicine and Rehabilitation, Harvard Medical School, Boston, MA, USA; [c]Spaulding Rehabilitation Hospital, Charlestown, MA, USA; [d]Home Base, A Red Sox Foundation and Massachusetts General Hospital Program, Boston, MA, USA; [e]Department of Psychology, Norwegian University of Science and Technology, Trondheim, Norway; [f]Department of Physical Medicine and Rehabilitation, St. Olavs Hospital University Hospital in Trondheim, Trondheim, Norway; [g]Department of Biological and Medical Psychology, University of Bergen, Bergen, Norway

ABSTRACT

The Cambridge Neuropsychological Test Automated Battery (CANTAB) is a battery of computerized neuropsychological tests commonly used in Europe in neurology and psychiatry studies, including clinical trials. The purpose of this study was to investigate test-retest reliability and to develop reliable change indices and regression-based change formulas for using the CANTAB in research and practice involving repeated measurement. A sample of 75 healthy adults completed nine CANTAB tests, assessing three domains (i.e., visual learning and memory, executive function, and visual attention) twice over a 3-month period. Wilcoxon signed-rank tests showed significant practice effects for 6 of 14 outcome measures with effect sizes ranging from negligible to medium (Hedge's *g*: .15–.40; Cliff's delta: .09–.39). The Spatial Working Memory test, Attention Switching Task, and Rapid Visual Processing test were the only tests with scores of adequate test-retest reliability. For all outcome measures, Pearson's and Spearman's correlation coefficients ranged from .39 to .79. The measurement error surrounding difference scores was large, thus requiring large changes in performance (i.e., 1–2 SDs) in order to interpret a change score as reliable. In the regression equations, test scores from initial testing significantly predicted retest scores for all outcome measures. Age was a significant predictor in several of the equations, while education was a significant predictor in only two of the equations. The adjusted R^2 values ranged between .19 and .67. The present study provides results enabling clinicians to make probabilistic statements about change in cognitive functions based on CANTAB test performances.

KEYWORDS
Neuropsychological testing; norms/normative studies; practice effects/ reliable change

Introduction

The Cambridge Neuropsychological Test Automated Battery (CANTAB) is a battery of computerized neuropsychological tests measuring multiple cognitive domains (Sahakian & Owen, 1992). It is commonly used in Europe in neurology (Ho et al., 2003; Williams-Gray, Foltynie, Brayne, Robbins, & Barker, 2007), psychiatry (Fried, Hirshfeld-Becker, Petty, Batchelder, & Biederman, 2015; Levaux et al., 2007), and neuropsychology research for studying diverse conditions, such as fetal alcohol spectrum disorders (Green et al., 2009), traumatic brain injury (TBI) (Sterr, Herron, Hayward, & Montaldi, 2006), Alzheimer's disease (O'Connell et al., 2004), affective disorders (Sweeney, Kmiec, & Kupfer, 2000), and schizophrenia (Hutton et al., 2004). It has been used in clinical trials involving treatment for depression (Falconer, Cleland, Fielding, & Reid, 2010), schizophrenia (Turner et al., 2004), and obsessive-compulsive disorder (Nielen & Den Boer, 2003). The CANTAB is being used in CENTER-TBI (Maas et al., 2015), a large European project that aims to improve the care for patients with TBI. CENTER-TBI is part of a larger global initiative called the International Initiative for Traumatic Brain Injury Research (InTBIR) with projects currently ongoing in Europe, the United States, and Canada.

The reliability of the CANTAB tests has not been thoroughly examined. Adequate reliability is a fundamental requirement for any test used in neuropsychology, regardless of its purpose (Crawford & Garthwaite, 2012). In classical test theory, reliability coefficients indicate the degree to which a test is free from measurement error, and consequently the confidence that clinicians place in test scores. Test-retest reliability concerns the

CONTACT Toril Skandsen toril.skandsen@ntnu.no Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, 7491 Trondheim, Norway; Department of Physical Medicine and Rehabilitation, St. Olavs Hospital University Hospital in Trondheim, 7006 Trondheim, Norway.

temporal stability of test scores and is of great importance for clinicians tracking change in cognitive functions over time. The Pearson's correlation coefficient (Pearson's *r*) is a value commonly used to estimate test-retest reliability.

When examining change in cognitive functions, clinicians must decide whether an individual's test score is meaningfully different from a score obtained in a previous evaluation, and not a reflection of measurement error. Several methods are available for this purpose (for a review, see Duff, 2012). Two of the most commonly used approaches involve using the reliable change methodology and standardized regression-based formulas. The reliable change methodology was used extensively in clinical psychology (Jacobson, Roberts, Berns, & McGlinchey, 1999) prior to being applied to clinical neuropsychology (Chelune, Naugle, Luders, Sedlak, & Awad, 1993; Heaton et al., 2001; Iverson, 2001; Temkin, Heaton, Grant, & Dikmen, 1999) and sports neuropsychology (Barr & McCrea, 2001; Hinton-Bayre, Geffen, Geffen, McFarland, & Friis, 1999; Iverson, Lovell, & Collins, 2003). This method involves the calculation of the Reliable Change Index (RCI), which indicates the probability that an observed difference between two test scores reflects measurement error. Because the traditional RCI-approach assumes no benefit of prior exposure to a test, a modification of the formula is recommended in the case of known practice effects (Chelune et al., 1993).

The standardized regression-based (SRB) approach involves using linear regression formulas to predict a retest score based on performance at initial testing (McSweeny, Naugle, Chelune, & Luders, 1993). This corrects for differential practice effects and regression toward the mean due to imperfect test reliability, as well as for variability in retest scores. Linear regression formulas are extendible to incorporate factors such as sample characteristics (e.g., age, gender, education) and testing schedule variables (e.g., test-retest interval) to predict retest scores. Regression-based change formulas have been used to investigate change in conditions such as epilepsy, TBI, and Parkinsons̀ disease (Duff, 2012).

To our knowledge, only a few studies have explored test-retest reliability of the CANTAB, three in older adults (Cacciamani et al., 2018; Goncalves, Pinho, & Simoes, 2016; Lowe & Rabbitt, 1998) and one in children (Syvaoja et al., 2015). Methodological differences (e.g., sample characteristics, administered tests, and test-retest interval) are evident between these studies. Nonetheless, a common finding is weak to moderate test-retest reliability for the majority of outcome measures, and only one of the studies used methods to evaluate change (Goncalves et al., 2016). There is a need for studies that compute reliable change statistics to refine the interpretation of the CANTAB in clinical practice. Therefore, the aim of the present study was to investigate the test-retest reliability of nine commonly used CANTAB tests across a three-month interval.

Methods

Participants

The participants were recruited as community controls in a large prospective cohort study on mild traumatic brain injury (MTBI) conducted as a collaboration between St. Olavs Hospital, Trondheim University Hospital and the Norwegian University of Science and Technology. The participants were matched at the group level regarding sex, age, and education to a sample of patients with MTBI. For practical reasons, they were recruited among the hospital and university staff, as well as families and friends of staff and patients with MTBI. Inclusion criteria were ages 16–59 years. Exclusion criteria were (a) non-residency in Norway or non-fluency in the Norwegian language; (b) ongoing severe psychiatric disease requiring treatment (e.g., bipolar disorder, severe depression), severe somatic disease, or substance abuse potentially making follow-up difficult; (c) history of complicated mild, moderate, or severe TBI or other preexisting neurological conditions with visible brain pathology or known cognitive deficits; and (d) MTBI in the last three months. One participant was excluded at the first visit due to a severe psychiatric disorder and one was excluded due to an unexpected MRI finding. Out of 81 participants who were assessed at the first visit, 75 returned for the second assessment and completed all tests. Only subjects assessed twice were included in the data analysis. The people not included in the data analysis were demographically similar to the overall sample, and we did not see a systematic reason for them to have not returned for the follow-up testing. Participants were not familiar with the CANTAB tests. None of the participants were diagnosed with Attention-Deficit/Hyperactivity Disorder, learning disability or used psychotropic medication. The participants (60% men) had a mean age of 32.21 years ($SD = 13.10$) with a mean level of education of 13.97 years ($SD = 2.44$, range: 10 to 18). The Regional Committees for Medical and Health Research Ethics (REC Central) approved this project and all participants gave informed consent.

Materials and procedures

All participants were assessed twice over a three-month period ($M = 3.10$ months, $SD = 0.37$, range: 1.92–4.32) with the same CANTAB tests, administered in the same order. This test-retest interval was used because this study is part of a larger observational cohort study investigating cognitive function following MTBI in adults, and testing three months after injury is a commonly used time point in MTBI research (Iverson, Karr, Gardner, Silverberg, & Terry, 2019). Well-trained research staff with bachelor or master level education in clinical psychology or neuroscience administrated the tests. All staff members were under supervision by a licensed clinical psychologist. Psychiatric disease was assessed with the Mini-International Neuropsychiatric Interview (Sheehan et al., 1998) administered by a clinical psychologist or medical doctor.

CANTAB

The CANTABeclipse^{TM} version 5.0.0 was used (Cambridge Cognition, 2012). Fourteen outcome measures from nine tests were included in the assessment procedure. Three tests were assumed to measure *visual learning and memory*

(Cambridge Cognition, 2012). The Paired Associates Learning (PAL) task presents participants with several white boxes that contain different patterns. Each pattern is subsequently revealed for one second and the participants must remember which box contains which pattern. The test was run in clinical mode and total errors adjusted for the number of trials was chosen as the outcome measure. A higher score is indicative of worse performance. The Pattern Recognition Memory (PRM) test presents participants with two different series of 12 patterns. Participants are then required to identify previously seen patterns among novel patterns immediately after the presentation (the first series) and after a 20-min delay (second series). The test was run in clinical mode and percent of correctly identified patterns for each trial was chosen as the outcome measure for each series. A higher score is indicative of better performance. The Spatial Recognition Memory (SRM) test presents the participants with a sequence of five white boxes appearing at various positions on the screen, and the participants must remember the screen placement for each of the boxes. The test was run in clinical mode and percent correct was chosen as the outcome measure. A higher score is indicative of better performance.

Four tests were assumed to measure *executive function* (Cambridge Cognition, 2012). In the Stockings of Cambridge (SOC) test, participants are shown two displays with three balls presented inside stockings, and the aim is to move the balls in the lower display such that it is identical to the arrangement of balls in the upper display. The test was run in clinical mode. The outcome measure was minimum number of possible moves, reflecting the sum of problems solved with the minimum number of possible moves. A higher score is indicative of better performance. In the Attention Switching Task (AST), participants are to determine the side or direction of an arrow on the screen. The arrow varies with respect to placement (right or left) and direction (right or left). The test was run in touch screen mode and three outcome measures were chosen. The first outcome measure, referred to as congruency cost, is the difference in mean response time in milliseconds on congruent (placement and direction are the same) and incongruent (placement and direction are *not* the same) trials. A positive score indicates that the participant is faster on congruent trials and a negative score indicates that the participant is faster on incongruent trials. The second outcome measure, switch cost, is the difference in mean response time in milliseconds on switch (where the current trial type and the previous trial type are the same, i.e., direction-direction or side-side) versus non-switch trials. A positive score indicates that the participant is faster on non-switch trials, and a negative score indicates that the participant is faster on switch trials. The third outcome measure is the percent of correct trials for both congruent and incongruent trials. A higher score (i.e., greater percent correct) is indicative of better performance. The Spatial Working Memory (SWM) test requires participants to search through boxes for a designated number of tokens. A token is never hidden in the same box twice; and to avoid errors, participants must remember where tokens originally appeared. The test was run in clinical mode and two outcome measures were chosen. The first outcome measure, between errors, is defined as the number of times the participant revisits a box in which a token has previously been found. A higher score is indicative of worse performance. The second outcome measure quantifies the effectiveness of the participant's strategy. This is a measure of the ability to follow a predetermined sequence beginning with a specific box and then to return to that box to start a new sequence once a blue token has been found. The minimum strategy score is 8 and the maximum is 56. A higher score is indicative of worse performance. The Spatial Span (SSP) test presents participants with multiple white boxes that change color one by one, and participants are asked to tap the boxes in the same order as they change color. The test was run in clinical mode and maximum span length (i.e., longest sequence) was chosen as the outcome measure. A higher score is indicative of better performance.

Two tests were assumed to measure *visual attention* (Cambridge Cognition, 2012). In the Rapid Visual Processing test (RVP), participants are presented numbers from 2 to 9 appearing inside a white box one at a time with a rate of 100 presentations per minute. The participants must press a button on a response box each time they see one of three target sequences (e.g., 2-4-6, 4-6-8, and 3-5-7). The test was run in clinical mode. A prime (A') is a measure of the ability to detect the target sequence and is the relationship between the probability of identifying a target sequence and the probability of identifying a non-target sequence. It ranges from .00 to 1.00 and a higher score is indicative of better performance. In the Reaction Time (RTI) test, the participant is to respond as fast as possible when a yellow dot is presented inside a circle (simple reaction time) and in one of five white circles (five-choice reaction time). The test was run in clinical mode and response time in milliseconds for each condition was chosen as the outcome measure. A higher score is indicative of worse performance.

Statistical analysis

All statistical analyses were conducted in R (R Core Team, 2017) using base R and relevant packages (*compute.es*: Del Re, 2014; *psych*: Revelle, 2016; *rsq*: Zhang, 2018). Raw test scores were used for all analyses because CANTAB only provides normalized scores for a small subset of all available tests and outcome measures. All participants successfully completed all CANTAB tests.

Several of the outcome measures violated the normality assumption with outliers present for most measures. Hence, differences in test scores between sessions were evaluated with Wilcoxon signed-rank test. Effect sizes were calculated using an unbiased Cohen's *d* (Hedges *g*: Hedges & Olkin, 1985) and Cliff's delta (Cliff, 1996). For Hedges *g*, an effect size $\leq$.20 was considered negligible, an effect size .21–.49 was considered small, an effect size .50–.79 was considered medium, and an effect sizes $\geq$.80 was considered large

Table 1. Test-retest data for the study sample.

Outcome measure	Time 1	Time 2	V	p	g	Δ	r (95% CI)	ρ (95% CI)
AST Congruency Cost	85.76 (70.23)	78.05 (67.32)	1,694	.156	.11	.07	.48 (.28–.64)	.47 (.28–.64)
AST Switch Cost	−130.08 (103.74)	−127.64 (96.34)	1,302	.518	.02	.02	.72 (.59–.82)	.73 (.59–.82)
AST % Correct	95.23 (6.40)	97.53 (4.83)	370	.000	.4	.39	.75 (.63–.84)	.52 (.63–.84)
PAL Total Errors Adj.	9.80 (12.15)	7.71 (14.86)	1,513	.009	.15	.18	.73 (.61–.82)	.59 (.61–.82)
PRM immediate	94.44 (9.22)	95.44 (8.02)	191.5	.381	.12	.05	.60 (.44–.73)	.46 (.44–.73)
PRM delayed	83.22 (14.66)	86.67 (13.63)	508.5	.043	.24	.14	.42 (.21–.59)	.40 (.21–.59)
RVP A'	.91 (.05)	.93 (.05)	570	.000	.37	.28	.75 (.63–.84)	.65 (.63–.84)
SRM % Correct	83.40 (10.85)	83.67 (11.52)	1,022	.906	.02	.03	.49 (.30–.65)	.46 (.30–.65)
SSP Span Length	6.76 (1.59)	6.84 (1.62)	417	.657	.05	.03	.69 (.55–.79)	.67 (.55–.79)
SWM Between Errors	16.11 (16.68)	13.00 (14.47)	1,448.5	.029	.2	.09	.71 (.58–.81)	.77 (.58–.81)
SWM Strategy	28.41 (6.80)	26.76 (6.73)	1,330	.006	.24	.15	.79 (.69–.86)	.79 (.69–.86)
RTI Simple Reaction Time	287.10 (36.84)	288.65 (37.89)	1,308	.538	.04	.02	.56 (.38–.70)	.47 (.38–.70)
RTI 5-choice Reaction Time	322.62 (42.35)	320.80 (41.97)	1,513	.501	.04	.03	.72 (.58–.81)	.70 (.58–.81)
SOC Min Moves	9.48 (2.02)	9.93 (1.54)	536	.072	.25	.11	.39 (.18–.57)	.43 (.18–.57)

Note. $N = 75$; for the columns Time 1 and Time 2, values represents raw score means and standard deviations (in parentheses). AST: Attention Switching Task; PAL: Paired Associates Learning; PRM: Pattern Recognition Memory; RVP: Rapid Visual Processing; SRM: Spatial Recognition Memory; SSP: Spatial Span; SWM: Spatial Working Memory; RTI: Reaction Time; SOC: Stockings of Cambridge; V: the sum of ranks assigned to the differences with positive sign; p: significance value for Wilcoxon signed rank test; g: Hedge's g; Δ: Cliff's delta; r: Pearson's correlation coefficient between time 1 and time 2 scores with 95% confidence interval (CI) in parentheses; ρ: Spearman's rank correlation coefficient between time 1 and time 2 scores with 95% confidence interval (CI) in parentheses.

Table 2. Mean difference score and reliable change estimates for CANTAB outcome measures.

					80% RCI			90% RCI		
					Unadjusted	Adjusted for practice effect		Unadjusted	Adjusted for practice effect	
Outcome measure	SEM_1	SEM_2	SE_{diff}	M_{diff}	Decline/Improvement	Decline	Improvement	Decline/Improvement	Decline	Improvement
AST Congruency Cost	50.80	48.69	70.36	−7.71	±90.06	−97.77	82.36	±115.39	−123.10	107.69
AST Switch Cost	54.59	50.70	74.50	2.45	±95.36	−92.91	97.81	±122.18	−119.73	124.63
AST % Correct	3.19	2.40	3.99	2.30	±5.11	−2.81	7.41	±6.54	−4.24	8.84
PAL Total Errors Adj.	6.26	7.65	9.89	−2.09	±12.65	10.56	−14.75	±16.21	14.12	−18.31
PRM Immediate	5.80	5.05	7.69	1.00	±9.84	−8.84	10.84	±12.61	−11.61	13.61
PRM Delayed	11.17	10.38	15.25	3.44	±19.52	−16.07	22.96	±25.00	−21.56	28.45
RVP A'	.02	.03	.03	.02	±0.04	−.03	.06	±0.06	−.04	.07
SRM % Correct	7.74	8.22	11.29	.27	±14.45	−14.18	14.71	±18.51	−18.25	18.78
SSP Span Length	.88	.90	1.26	.08	±1.61	−1.53	1.69	±2.07	−1.99	2.15
SWM Between Errors	8.98	7.79	11.88	−3.11	±15.21	12.10	−18.31	±19.49	16.38	−22.59
SWM Strategy	3.10	3.07	4.36	−1.65	±5.58	3.93	−7.24	±7.16	5.50	−8.81
RTI Simple Reaction Time	24.55	25.25	35.22	1.55	±45.08	46.63	−43.53	±57.76	59.31	−56.21
RTI 5-choice Reaction Time	22.60	22.40	31.82	−1.83	±40.73	38.90	−42.56	±52.19	50.36	−54.01
SOC Min Moves	1.58	1.20	1.98	.45	±2.53	−2.08	2.99	±3.24	−2.79	3.70

Note. $N = 75$; AST: Attention Switching Task; PAL: Paired Associates Learning; PRM: Pattern Recognition Memory; RVP: Rapid Visual Processing; SRM: Spatial Recognition Memory; SSP: Spatial Span; SWM: Spatial Working Memory; RTI: Reaction Time; SOC: Stockings of Cambridge; SEM: Standard error of measurement for time 1 and time 2; SE_{diff}: Standard error of difference; M_{diff}: Mean difference score.

(Cohen, 1992). For Cliff's delta, an effect size ≤.15 was considered negligible, an effect size .16–.33 was considered small, an effect size .34–.47 was considered medium, and an effect size ≥.47 was considered large (Romano, Kromrey, Coraggio, & Skowronek, 2006). Test-retest reliability was calculated with both Pearson product-moment correlation coefficients (r) and Spearman's rank correlation coefficients (ρ). The level for acceptable test-retest reliability was defined as ≥.75, in accordance with previously recommended reliability levels using the CANTAB (Lowe & Rabbitt, 1998). The standard error of measurement (SEM) for each session was calculated as follows:

$$SEM = SD\sqrt{1 - r}$$

where SD is the standard deviation from the session and r is the test-retest Pearson's product-moment correlation coefficient. RCIs were calculated based on the standard error of difference (SE_{diff}), calculated according to Iverson (2001):

$$SE_{diff} = \sqrt{SEM_1^2 + SEM_2^2}$$

where SEM_1 and SEM_2 are the SEM from the first and second sessions, respectively. Each confidence interval (CI) was calculated by multiplying the SE_{diff} with a specific z-score (i.e., 80% CI: $z = 1.28$ and 90% CI: $z = 1.64$). For all outcome measures, the mean practice effects [i.e., Mean Time 2 ($T2$) − Mean Time 1 ($T1$)] were added to the lower and upper bounds of the CI for the RCI (Chelune et al., 1993).

Regression-based change formulas (SRBs) using multiple regression equations were developed, in which scores from the first session ($T1$) were placed into a linear regression equation with scores from the second session ($T2$) as the dependent variable and age, gender, and education as covariates. Insignificant predictors ($p > .05$) were removed with stepwise regression using backwards selection. Predictors were removed in the following order: sex, education, and age. Of note, for all models, the mean of the residuals was approximately zero and equal residual variance was present. Variance inflation factors were low (<2) for all covariates in all models. Durbina-Watson test did not show autocorrelation of residuals and all covariates and residuals were uncorrelated. However, deviations from normality for the

Table 3. Regression equations for CANTAB outcome measures.

Outcome measure	F(df)	R^2	SEE	Predicted $T2$	Partial R^2
AST Congruency Cost	13.52 (2,72)	.25	58.19	4.74 + ($T1$*.42) + (age*1.12)	$T1$ = .19; Age = .05
AST Switch Cost	54.27 (2,72)	.59	61.68	19.76 + ($T1$*.56) + (age*−2.27)	$T1$ = .42; Age = .15
AST % Correct	95.04 (1,73)	.56	3.20	43.51 + ($T1$*.57)	$T1$ = .56
PAL Total Errors Adj.	85.64 (1,73)	.53	10.15	−1.10 + ($T1$*.90)	$T1$ = .53
PRM Immediate	42.03 (1,73)	.36	6.43	45.79 + ($T1$*.53)	$T1$ = .36
PRM Delayed	11.94 (2,72)	.23	11.98	68.47 + ($T1$*.33) −(age* .29)	$T1$ = .13; Age = .08
RVP A'	94.84 (1,73)	.56	.03	.17 + ($T1$*.83)	$T1$ = .56
SRM % Correct	14.31 (3,71)	.35	9.28	50.79 + ($T1$*.32) −(age*.31) + (edu*1.15)	$T1$ = .08; Age = .11; Education = .07
SSP Span Length	38.42 (2,72)	.50	1.14	4.14 + ($T1$*.55) −(age*.03)	$T1$ = .26; Age = .06
SWM Between Errors	43.59 (2,72)	.54	9.86	−4.39 + ($T1$*.47) + (age*.29)	$T1$ = .27; Age = .07
SWM Strategy	67.57 (2,72)	.64	4.02	3.97 + ($T1$*.69) + (age*.09)	$T1$ = .50; Age = .05
RTI Simple Reaction Time	25.38 (2,72)	.40	29.42	114.88 + ($T1$*.49) + (age*.96)	$T1$ = .26; Age = .14
RTI 5-choice Reaction Time	50.33 (3,71)	.67	24.23	137.24 + ($T1$*.56) + (age*1.32) −(edu*−2.99)	$T1$ = .46; Age = .31; Education = .07
SOC Min Moves	9.53 (2,72)	.19	1.39	9.10 + ($T1$*.20) −(age*.03)	$T1$ = .05; Age = .05

Note. $N = 75$; All F-tests are significant at $p < .001$. AST: Attention Switching Task; PAL: Paired Associates Learning; PRM: Pattern Recognition Memory; RVP: Rapid Visual Processing; SRM: Spatial Recognition Memory; SSP: Spatial Span; SWM: Spatial Working Memory; RTI: Reaction Time; SOC: Stockings of Cambridge; R^2: Adjusted R^2; SEE: Standard error of the estimate.

Table 4. Interpreting change on the CANTAB based on the natural distribution of difference scores (Time 2–Time 1).

	Decline		Improvement	
Outcome measures	Very uncommon 5%	Uncommon 10%	Very uncommon 5%	Uncommon 10%
AST Congruency Cost	−94.85	−89.08	117.57	37.49
AST Switch Cost	−129.19	−91.81	108.80	71.52
AST % Correct	−2.06	−1.25	9.25	3.75
PAL Total Errors Adj.	7.00	2.00	−21.20	−10.00
PRM Immediate	−8.33	−8.33	10.83	8.33
PRM Delayed	−19.17	−16.67	25.00	16.67
RVP A'	−.03	−.03	.09	.04
SRM % Correct	−15.00	−10.00	20.00	10.00
SSP Span Length	−2.00	−1.00	2.30	1.00
SWM Between Errors	11.20	3.00	−23.30	−16.60
SWM Strategy	3.00	1.00	−9.90	−7.60
RTI Simple Reaction Time	56.68	27.73	−62.18	−46.55
RTI 5-choice Reaction Time	47.98	21.65	−55.03	−34.73
SOC Min Moves	−3.00	−2.00	3.30	2.00

Note. $N = 75$; AST: Attention Switching Task; PAL: Paired Associates Learning; PRM: Pattern Recognition Memory; RVP: Rapid Visual Processing; SRM: Spatial Recognition Memory; SSP: Spatial Span; SWM: Spatial Working Memory; RTI: Reaction Time; SOC: Stockings of Cambridge.

residuals, as well as outliers and influential cases were seen in several models. Predicted $T2$ scores were subtracted from the obtained $T2$ scores and divided by the standard error of the estimate (SEE). The calculation of the SRB results in a z-score. A z-score of ± 1.65 was chosen as the demarcation point for reliable change, indicating that 10% (i.e., 5% at each tail of the curve) of change scores will fall beyond this cutoff.

In addition to the RCI and SRB approaches to determining reliable change, the natural distribution of change scores ($T2 - T1$) for determining decline or improvement on the CANTAB is presented in Table 4. Unlike the RCI and SRB methods, this approach makes no assumption about normality of the data, rather providing raw values of change scores that fell below or above a specific cumulative percentage of our sample.

Results

Mean scores for the first and second sessions are provided for each outcome measure in Table 1. Statistically significant differences ($\alpha = .05$) in test scores between sessions were seen for AST percent correct, RVP A', SWM strategy, PAL total errors adjusted, SWM between error, and PRM delayed recall. Improved performance from session 1 to session 2 was seen on all measures, with effect sizes ranging from negligible to medium. The largest practice effects were seen for AST percent correct ($g = .40$, delta = .39) and RVP A' ($g = .37$, delta = .28). Pearson's product-moment correlation coefficients above the cutoff level for acceptability of $\geq .75$ (Lowe & Rabbitt, 1998) were obtained only for SWM strategy, AST percent correct, and RVP A'. Only SWM strategy and SWM between errors had a Spearman's rank correlation coefficient $>.75$.

Table 2 shows mean difference scores, $SEMs$ for each session, SE_{diff}, and RCIs with and without adjustment for practice effects. Large changes in test scores were required for reliable change for all outcome measures, ranging from one SD of the $T1$ score for AST percent correct to nearly two SDs of the $T1$ score for SRM percent correct (See Table 1). Table 3 shows the results from the regression equations. The F, R^2, SEE, unstandardized beta weights, and the constant for each outcome measure are provided in Table 3. All *F*-tests were significant ($p < .001$), indicating that the regression models provided a better fit than the intercept-only model. Age and education were only significant predictors in some of the models. Across CANTAB tests, the models accounted for between 19% and 67% of the variance (adjusted R^2). Partial adjusted R^2 values are provided for all significant predictors. Table 4 provides change scores at the 5th, 10th, 90th, and 95th percentiles of the natural distribution of change scores for our sample.

Discussion

This study presents three-month test-retest data, as well as reliable change indices and regression-based formulas for several outcome measures from the CANTAB, thereby extending the current literature and facilitating the use of the CANTAB in clinical practice. Practice effects were seen for several outcome measures, a finding consistent with existing literature across tests from different cognitive domains (Calamia, Markon, & Tranel, 2012). Acceptable test-retest correlations of $r \geq .75$ (Lowe & Rabbitt, 1998) were obtained for only SWM between errors, AST percent correct, and RVP A'; and only SWM strategy and SWM between errors had Spearman's correlation coefficients of $\rho > .75$. Thus, the findings are consistent with previous studies (Cacciamani et al., 2018; Goncalves et al., 2016; Lowe & Rabbitt, 1998; Syvaoja et al., 2015), demonstrating low to medium reliability coefficients for the majority of CANTAB tests. Consistent with prior research studies in adults (Cacciamani et al., 2018; Goncalves et al., 2016; Lowe & Rabbitt, 1998), inadequate test-retest reliability was demonstrated for PAL total errors adjusted, PRM delayed recall, SRM percent correct, SSP span length, RTI simple and five-choice reaction time, and SOC minimum number of possible moves. Our finding of adequate test-retest reliability for SWM between errors and strategy, as well as RVP A', is somewhat surprising, and is not consistent with prior research studies (Cacciamani et al., 2018; Goncalves et al., 2016). However, this inconsistency may be explained by the fact that these studies have included older adults, some with cognitive impairment, which is known to affect test-retest reliability (Calamia, Markon, & Tranel, 2013; Duff, 2012).

Low test-retest reliability is common in neuropsychology, and the reliability coefficients obtained for the CANTAB are similar to those associated with commonly used neuropsychological test batteries, such as the Delis-Kaplan Executive Function System (D-KEFS; Delis, Kaplan & Kramer, 2001), Wechsler Memory Scale, Third Edition (WMS-III; Wechsler, 1997), and Neuropsychological Assessment Battery (NAB; Stern & White, 2003). A common theme in psychometric research is that memory and executive functions are difficult to assess in a reliable manner (Calamia et al., 2013; Strauss, Sherman, & Spreen, 2006). Some authors have suggested that excellent tests of executive functions will inevitably have low temporal stability because these tests, by design, require novelty (Rabbitt, Lowe, & Shilling, 2001). Furthermore, it is reasonable to assume that successful performance on memory tests is, at least partially, dependent on executive functions, such as working memory and strategic approaches to learning; thus, affecting the temporal stability of memory tests. In addition, the memory tests used in our study exposed participants to the same information twice, thusly affecting test-retest reliability. Regardless, low reliability limits a test's utility for diagnostic purposes and its usefulness for detecting change over time (Strauss et al., 2006).

We developed reliable change indices and regression-based change formulas for 14 outcome measures from 9 CANTAB tests. The measurement error surrounding difference scores indicated that relatively large changes in performance were needed to interpret a change as reliable, ranging from one SD of the *T*1 score for AST percent correct to nearly two SDs of the *T*1 score for SRM percent correct. Consistent with previous research on healthy adults (Attix et al., 2009; Duff et al., 2010; Duff et al., 2004; Duff et al., 2005; Sánchez-Benavides et al., 2016; Temkin et al., 1999), test scores from initial testing significantly predicted retest scores for all outcome measures. Furthermore, age was a significant predictor in many tests across all neuropsychological domains, including AST congruency cost and switch cost, PRM delayed recall, SRM percent correct, SSP span length, SWM between errors and strategy, RTI simple and five-choice reaction time, and SOC minimum number of possible moves. Education contributed significantly only in one test of visual memory (SRM percent correct) and one test of attention (RTI five-choice reaction time). These findings are inconsistent with the study by Goncalves et al. (2016), which found the best fit when excluding age and education from the regression models. However, these differences may be explained by the small sample size utilized by Goncalves et al. (2016), as studies with larger samples consistently have shown effects of both age and education on tests across multiple cognitive domains (Duff, 2012). Our finding of an adjusted R^2 ranging between .19 and .67 indicated that additional variance in retest scores was unexplained by the different regression models. However, proportions of explained variance for the CANTAB tests were similar to findings from research on healthy adults using other neuropsychological tests measuring a broad range of cognitive domains (Attix et al., 2009; Duff et al., 2010; Duff et al., 2004; Duff et al., 2005; Sánchez-Benavides et al., 2016; Temkin et al., 1999). To illustrate the clinical use of reliable change indices, regression-based change formulas, and cutoffs from our observed change score distribution, we present a fictional case example in the Appendix.

Currently, consensus is lacking on the best method for evaluating reliable change (Hinton-Bayre, 2016). We chose to supplement the more traditional RCI approach with the SRB methodology because it takes into account several elements of variability (Chelune et al., 1993; Iverson, 2001). However, when comparing different methods, research has often produced similar results (Barr & McCrea, 2001; Heaton et al., 2001; Hinton-Bayre, 2012; Maassen, Bossema, & Brand, 2009). In addition, considered non-normality of our data, we provided raw percentiles from our observed distributions, through which a clinician could make a comparisons as to where a change score would fall compared to others within our sample. By using the data provided in this paper, clinicians have the opportunity to choose the methods most suited for their particular clinical situation, whether they wish to adjust for practice effects, consider age and education, or make normality assumptions in their determination of change.

Although our results have applications for use of the CANTAB, our study design does include limitations that researchers and clinicians should consider when translating our findings into their research designs or clinical approach. Reliability coefficients are influenced by many different factors such as the age and health of participants, as well as the length of the test-retest interval (Calamia, Markon, & Tranel, 2012). Since we applied a three-month test-retest interval and an age range from 16 to 60 years, our results may not be generalizable to assessments with longer or shorter test-retest intervals, or to patients and participants outside the age range of our sample (i.e., pediatric or geriatric populations). The three-month interval was chosen as the study was part of a larger study investigating cognitive functioning following MTBI in adults. A shorter test-retest interval may be more appropriate for some tests that may be re-administered multiple times over the course of recovery following an MTBI; and a longer test-retest interval may be more appropriate for other tests, if they are more often re-administered with longer intervals in clinical practice. However, the magnitude of the test-retest correlation has been shown to decrease with increasing time interval (Duff, 2012).

Furthermore, the sample size in our study is relatively small, which may affect the accuracy of our results. However, the sample size is comparable to other studies on the reliability of the CANTAB (Cacciamani et al., 2018; Goncalves et al., 2016; Syvaoja et al., 2015). We did not recruit participants directly from the community but used a convenience sampling approach to recruit hospital and university staff, as well as families and friends of staff and patients with MTBI. The mean education level in our sample was also fairly high (i.e., 14 years), which limits the application of our findings to participant of lower education levels. Thus, the generalizability of our results would be informed through replication with larger and more diverse samples of participants and through further studies on the CANTAB using different test-retest intervals. Another limitation in our study design is that we did not administer performance validity tests. However, none of the participants were involved in litigation and there were no other known external incentives.

Of note, our findings evidence significant limitations as to the reliability of CANTAB test scores, and we made judgements about the inadequacy of test-retest reliability based on a selected cutoff of $\geq$.75. Although we selected this cutoff, no universally accepted cutoff exists for defining adequate reliability. In the present study, we chose to describe reliabilities according to the labels used by Lowe and Rabbitt (1998), but if we had chosen a lower cutoff for adequate reliability, such as .70 (Strauss et al., 2006), the outcome measures of AST switch cost, PAL total errors adjusted, and RTI five-choice reaction time would have been classified as acceptable. However, the calculations used to determine reliable change would not change.

A final limitation pertains to non-normality of our data. The calculation of RCIs and regression formulas for determining reliable change make certain assumptions concerning the properties of our data. We chose to approach the determination of change in three ways, including a simple description of cutoffs in our distribution that makes no assumption of normality. There are many scores that may be administered repeatedly in research or clinical practice, which often, by design, present with non-normal distributions, because they either occur infrequently (e.g., errors), or have lower bound limits to performance (e.g., reaction time). As computerized tasks such as the CANTAB become more common in clinical practice, researchers may need to develop more sophisticated methods for interpreting individual change on tests with non-normal distributions that consider important aspects related to test performance (e.g., retest effects, age, education, etc.). Furthermore, neuropsychologists frequently evaluate patients on more than two time points, and it is unlikely that the results from this study can be used to investigate change between a second and a third time point. Future research should investigate change over multiple assessment sessions using the methods from this paper, as well as utilizing other statistical methods such as latent curve modeling (Duff, 2012).

In summary, the results of this study have implications for those who use the CANTAB in research and clinical practice. Practice effects were seen for several outcome measures, with AST percent correct and RVP A' demonstrating the largest effect sizes. Acceptable levels of test-retest reliability were only seen SWM between errors and strategy, AST percent correct, and RVP A'. Thus, the probable range of measurement error surrounding most test-retest difference scores is large for the CANTAB, meaning that large changes in performance are needed before a clinician or researcher can conclude with confidence that the observed change is not due to measurement error. The results from this paper allow neuropsychologists to consider these factors and make probabilistic statements about change using reliable change indices, standardized regression equations, and the distribution of change scores.

ORCID

Rune H. Karlsen http://orcid.org/0000-0002-3435-9156
Justin E. Karr http://orcid.org/0000-0003-3653-332X
Grant L. Iverson http://orcid.org/0000-0001-7348-9570
Toril Skandsen http://orcid.org/0000-0001-5495-9338

References

Attix, D. K., Story, T. J., Chelune, G. J., Ball, J. D., Stutts, M. L., Hart, R. P., & Barth, J. T. (2009). The prediction of change: Normative neuropsychological trajectories. *The Clinical Neuropsychologist*, *23*(1), 21–38. doi:10.1080/13854040801945078

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. 11575591 doi:10.1017/S1355617701766052

Cacciamani, F., Salvadori, N., Eusebi, P., Lisetti, V., Luchetti, E., Calabresi, P., & Parnetti, L. (2018). Evidence of practice effect in CANTAB spatial working memory test in a cohort of patients with mild cognitive impairment. *Applied Neuropsychology: Adult*, *25*(3), 237–248. doi:10.1080/23279095.2017.1286346

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist*, *26*(4), 543–570. doi:10.1080/13854046.2012.680913

Calamia, M., Markon, K., & Tranel, D. (2013). The robust reliability of neuropsychological measures: Meta-analyses of test-retest correlations. *The Clinical Neuropsychologist*, *27*(7), 1077–1105. doi:10.1080/13854046.2013.809795

Cambridge Cognition. (2012). *CANTABeclipse 5: Test administration guide*. Cambridge, UK: Cambridge Cognition.

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. doi:10.1037/0894-4105.7.1.41

Cliff, N. (1996). *Ordinal methods for behavioral data analysis*. New York, NY: Psychology Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. doi:10.1037/0033-2909.112.1.155

Crawford, J. R., & Garthwaite, P. H. (2012). Single-case research in neuropsychology: A comparison of five forms of t-test for comparing a case to controls. *Cortex*, *48*(8), 1009–1016. doi:10.1016/j.cortex.2011.06.021

Del Re, A. C. (2014). *Compute.es: Compute effect sizes*. Available from https://cran.r-project.org/web/packages/compute.es/

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan executive function system*. San Antonio, TX: The Psychological Corporation.

Duff, K. (2012). Evidence-based indicators of neuropsychological change in the individual patient: Relevant concepts and methods. *Archives of Clinical Neuropsychology*, *27*(3), 248–261. doi:10.1093/arclin/acr120

Duff, K., Beglinger, L. J., Moser, D. J., Paulsen, J. S., Schultz, S. K., & Arndt, S. (2010). Predicting cognitive change in older adults: The relative contribution of practice effects. *Archives of Clinical Neuropsychology*, *25*(2), 81–88. doi:10.1093/arclin/acp105

Duff, K., Schoenberg, M. R., Patton, D., Mold, J., Scott, J. G., & Adams, R. L. (2004). Predicting change with the RBANS in a community dwelling elderly sample. *Journal of the International Neuropsychological Society*, *10*(6), 828–834. doi:10.1017/S1355617704106048

Duff, K., Schoenberg, M., Patton, D., Paulsen, J., Bayless, J., Mold, J., … Adams, R. (2005). Regression-based formulas for predicting change in RBANS subtests with older adults. *Archives of Clinical Neuropsychology*, *20*(3), 281–290. doi:10.1016/j.acn.2004.07.007

Falconer, D. W., Cleland, J., Fielding, S., & Reid, I. C. (2010). Using the Cambridge Neuropsychological Test Automated Battery (CANTAB) to assess the cognitive impact of electroconvulsive therapy on visual and visuospatial memory. *Psychological Medicine*, *40*(6), 1017–1025. doi:10.1017/S0033291709991243

Fried, R., Hirshfeld-Becker, D., Petty, C., Batchelder, H., & Biederman, J. (2015). How informative is the CANTAB to assess executive functioning in children with ADHD? A controlled study. *Journal of Attention Disorders*, *19*(6), 468–475. doi:10.1177/1087054712457038

Goncalves, M. M., Pinho, M. S., & Simoes, M. R. (2016). Test-retest reliability analysis of the Cambridge Neuropsychological Automated Tests for the assessment of dementia in older people living in retirement homes. *Applied Neuropsychology: Adult*, *23*(4), 251–263. doi:10.1080/23279095.2015.1053889

Green, C. R., Mihic, A. M., Nikkel, S. M., Stade, B. C., Rasmussen, C., Munoz, D. P., & Reynolds, J. N. (2009). Executive function deficits in children with fetal alcohol spectrum disorders (FASD) measured using the Cambridge Neuropsychological Tests Automated Battery (CANTAB). *Journal of Child Psychology and Psychiatry*, *50*(6), 688–697. doi:10.1111/j.1469-7610.2008.01990.x

Heaton, R. K., Temkin, N., Dikmen, S., Avitable, N., Taylor, M. J., Marcotte, T. D., & Grant, I. (2001). Detecting change: A comparison of three neuropsychological methods, using normal and clinical samples. *Archives of Clinical Neuropsychology*, *16*(1), 75–91. doi:10.1093/arclin/16.1.75

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hinton-Bayre, A. D. (2012). Choice of reliable change model can alter decisions regarding neuropsychological impairment after sports-related concussion. *Clinical Journal of Sport Medicine*, *22*(2), 105–108. doi:10.1097/JSM.0b013e318248a526

Hinton-Bayre, A. D. (2016). Clarifying discrepancies in responsiveness between reliable change indices. *Archives of Clinical Neuropsychology*, *31*(7), 754–768. doi:10.1093/arclin/acw064

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. doi:10.1076/jcen.21.1.70.945

Ho, A.K., Sahakian, B.J., Brown, R.G., Barker, R.A., Hodges, J.R., Ane, M.-N., … Bodner, T. (2003). Profile of cognitive progression in early Huntington's disease. *Neurology*, *61*(12), 1702–1706. doi:10.1212/01.WNL.0000098878.47789.BD

Hutton, S. B., Huddy, V., Barnes, T. R. E., Robbins, T. W., Crawford, T. J., Kennard, C., & Joyce, E. M. (2004). The relationship between antisaccades, smooth pursuit, and executive dysfunction in first-episode schizophrenia. *Biological Psychiatry*, *56*(8), 553–559. doi:10.1016/j.biopsych.2004.07.002

Iverson, G. L. (2001). Interpreting change on the WAIS-III/WMS-III in clinical samples. *Archives of Clinical Neuropsychology*, *16*(2), 183–191. doi:10.1093/arclin/16.2.183

Iverson, G. L., Karr, J. E., Gardner, A. J., Silverberg, N. D., & Terry, D. P. (2019). Results of scoping review do not support mild traumatic brain injury being associated with a high incidence of chronic cognitive impairment: Commentary on McInnes et al. 2017. *PloS One*, *14*(9), e0218997. doi:10.1371/journal.pone.0218997

Iverson, G. L., Lovell, M. R., & Collins, M. W. (2003). Interpreting change on ImPACT following sport concussion. *The Clinical Neuropsychologist*, *17*(4), 460–467. doi:10.1076/clin.17.4.460.27934

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology*, *67*(3), 300–307. doi:10.1037/0022-006X.67.3.300

Levaux, M. N., Potvin, S., Sepehry, A. A., Sablier, J., Mendrek, A., & Stip, E. (2007). Computerized assessment of cognition in schizophrenia: Promises and pitfalls of CANTAB. *European Psychiatry*, *22*(2), 104–115. doi:10.1016/j.eurpsy.2006.11.004

Lowe, C., & Rabbitt, P. (1998). Test/re-test reliability of the CANTAB and ISPOCD neuropsychological batteries: Theoretical and practical issues. Cambridge Neuropsychological Test Automated Battery. International Study of Post-Operative Cognitive Dysfunction. *Neuropsychologia*, *36*(9), 915–923. 9740364 doi:10.1016/S0028-3932%2898%2900036-0

Maas, A. I. R., Menon, D. K., Steyerberg, E. W., Citerio, G., Lecky, F., Manley, G. T., … Sorgner, A. (2015). Collaborative European NeuroTrauma Effectiveness Research in Traumatic Brain Injury (CENTER-TBI): A prospective longitudinal observational study. *Neurosurgery*, *76*(1), 67–80. doi:10.1227/NEU.0000000000000575

Maassen, G. H., Bossema, E., & Brand, N. (2009). Reliable change and practice effects: Outcomes of various indices compared. *Journal of Clinical and Experimental Neuropsychology*, *31*(3), 339–352. doi:10.1080/13803390802169059

McSweeny, 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. [Peer Reviewed]. *Clinical Neuropsychologist*, *7*(3), 300–312. doi:10.1080/13854049308401901

Nielen, M. M., & Den Boer, J. A. (2003). Neuropsychological performance of OCD patients before and after treatment with fluoxetine:

Evidence for persistent cognitive deficits. *Psychological Medicine*, *33*(5), 917–925. doi:10.1017/S0033291703007682

O'Connell, H., Coen, R., Kidd, N., Warsi, M., Chin, A. V., & Lawlor, B. A. (2004). Early detection of Alzheimer's disease (AD) using the CANTAB paired Associates Learning Test. *International Journal of Geriatric Psychiatry*, *19*(12), 1207–1208. doi:10.1002/gps.1180

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Available from http://www.R-project.org/

Rabbitt, P., Lowe, C., & Shilling, V. (2001). Frontal tests and models for cognitive ageing. *The European Journal of Cognitive Psychology*, *13*(1–2), 5–28. doi:10.1080/09541440042000197

Revelle, W. (2016). *psych: Procedures for personality and psychological research*. Available from https://CRAN.R-project.org/package=psych.

Romano, J., Kromrey, J. D., Coraggio, J., & Skowronek, J. (2006). Should we really be using t-test and Cohen's d for evaluating group differences on the NSSE and other surveys. Paper presented at the Annual meeting of the Florida association of institutional research.

Sahakian, B. J., & Owen, A. M. (1992). Computerized assessment in neuropsychiatry using CANTAB: Discussion paper. *Journal of the Royal Society of Medicine*, *85*(7), 399–402.

Sánchez-Benavides, G., Peña-Casanova, J., Casals-Coll, M., Gramunt, N., Manero, R. M., Puig-Pijoan, A., Aguilar, M., & Ribas, R. (2016). One-year reference norms of cognitive change in Spanish old adults: Data from the NEURONORMA sample. *Archives of Clinical Neuropsychology*, *31*(4), 378–388. doi:10.1093/arclin/acw018

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., & Weiller, E. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of Clinical Psychiatry*, *59* (Suppl 20), 22–33. quiz 34–57.

Stern, R. A., & White, T. (2003). *Neuropsychological assessment battery: Administration, scoring, and interpretation manual*. Lutz, FL: Psychological Assessment Resources

Sterr, A., Herron, K. A., Hayward, C., & Montaldi, D. (2006). Are mild head injuries as mild as we think? Neurobehavioral concomitants of chronic post-concussion syndrome. *BMC Neurology*, *6*(1), 7. doi:10.1186/1471-2377-6-7

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary* (3rd ed.). New York, NY: Oxford University Press.

Sweeney, J. A., Kmiec, J. A., & Kupfer, D. J. (2000). Neuropsychologic impairments in bipolar and unipolar mood disorders on the CANTAB neurocognitive battery. *Biological Psychiatry*, *48*(7), 674–684. doi:10.1016/S0006-3223(00)00910-0

Syvaoja, H. J., Tammelin, T. H., Ahonen, T., Rasanen, P., Tolvanen, A., & Kankaanpaa, A. (2015). Internal consistency and stability of the CANTAB neuropsychological test battery in children. *Psychological Assessment*, *27*(2), 698–709. doi:10.1037/a0038485

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. doi:10.1017/S1355617799544068

Turner, D. C., Clark, L., Pomarol-Clotet, E., McKenna, P., Robbins, T. W., & Sahakian, B. J. (2004). Modafinil improves cognition and attentional set shifting in patients with chronic schizophrenia. *Neuropsychopharmacology*, *29*(7), 1363–1373. doi:10.1038/sj.npp.1300457

Wechsler, D. (1997). *Wechsler Memory Scale*–Third Edition. San Antonio, TX: Psychological Corporation.

Williams-Gray, C. H., Foltynie, T., Brayne, C. E., Robbins, T. W., & Barker, R. A. (2007). Evolution of cognitive dysfunction in an incident Parkinson's disease cohort. *Brain*, *130*(7), 1787–1798. doi:10.1093/brain/awm111

Zhang, D. (2018). *rsq: R-squared and related measures*. Available from https://CRAN.R-project.org/package=rsq

Appendix

Case example

To illustrate the clinical use of reliable change indices, regression-based change formulas, and cutoffs from our observed change score distribution, we present a fictional case example of a 25-year-old man with 15 years of education who had a traumatic brain injury of moderate severity. The patient is tested three and 6 months following injury with RTI five-choice reaction time. Mean reaction time was 430 ms at 3-month testing and 370 ms at 6-month testing.

Reliable change index

This 60 ms decrease in reaction time is above the cutoff of 54 ms from the RCI 90% CI after adjusting for practice effects (see Table 2), indicating that a reliable change has occurred. In the absence of a reference table, or if the clinician is interested in using a different confidence interval, the calculation can also be done manually with the following formula:

$$RCI = \frac{(T2 - T1) - (M2 - M1)}{SE_{diff}}$$

where the mean practice effect ($M2$–$M1$ = −1.83) is subtracted from the difference score for the individual ($T2$–$T1 = 60$) and divided by the standard error of the difference (31.82). This calculation results in a z-value of −1.8, which is below the −1.65 demarcation point, approximately at the third percentile. If a more stringent criterion of $z \pm 1.96$ (i.e., 95% confidence interval) is used, the change is not interpreted as reliable.

Regression based change formula

Using this approach, the first step is to calculate the predicted retest score:

$$T'_2 = b_1 T_1 + b_2 \text{Age} + \ b_3 \text{Education} + \ c$$

where T'_2 is the predicted retest score, b_1 is the regression slope for initial testing, $T1$ is the score from initial testing, and c the regression intercept. As age and education were significant predictors in the model for the RTI five-choice reaction time (see Table 3), the regression slope for age (b_2) and education in years (b_3) is included in the equation.

Using the information provided in the example (observed $T1$ and $T2$ test scores, age, and education) in combination with the data in Table 3 (regression slopes and the intercept), the predicted retest score for RTI five-choice reaction time would be

$$T'_2 = \ .56 \ \times \ 430 + 1.32 \ \times \ 25 - 3.00 \ \times \ 15 + 137 = \ 366$$

The predicted retest score is then tested as follows

$$RCI_{SRB} = \frac{T_2 - T'_2}{SEE}$$

where SEE is the standard error of the estimate of the regression equation. The resulting value [(370 − 366)/24.23 = 0.17] is then compared with a normal distribution table, and ± 1.64 is used as the cutoff for defining reliable change. The value is within the ±1.64 interval, indicating that the decrease in mean five-choice of 60 ms does not reflect a reliable change.

Absolute differences based on the distribution of change scores

A final method for evaluating change would use cutoffs from the distribution of change scores presented in Table 4. One can see that the 60 ms decrease in reaction time is below the fifth percentile, indicating that the improvement has a low likelihood of happening by chance, because less than 5% of our observed sample obtained such a change score.