

Test-retest reliability and reliable change index of the Philips IntelliSpace Cognition digital test battery

Laura Klaming, Mandy Spaltman, Stefan Vermeent, Gijs van Elswijk, Justin B. Miller & Ben Schmand

To cite this article: Laura Klaming, Mandy Spaltman, Stefan Vermeent, Gijs van Elswijk, Justin B. Miller & Ben Schmand (15 Feb 2024): Test-retest reliability and reliable change index of the Philips IntelliSpace Cognition digital test battery, *The Clinical Neuropsychologist*, DOI: [10.1080/13854046.2024.2315747](https://doi.org/10.1080/13854046.2024.2315747)

To link to this article: <https://doi.org/10.1080/13854046.2024.2315747>



Published online: 15 Feb 2024.



Submit your article to this journal [↗](#)




View related articles [↗](#)



View Crossmark data [↗](#)



Test–retest reliability and reliable change index of the Philips IntelliSpace Cognition digital test battery

Laura Klaming^a, Mandy Spaltman^a, Stefan Vermeent^a , Gijs van Elswijk^a, Justin B. Miller^b  and Ben Schmand^a

^aDigital Cognitive Diagnostics, Philips Healthcare, Eindhoven, The Netherlands; ^bCleveland Clinic Lou Ruvo Center for Brain Health, Las Vegas, NV, USA

ABSTRACT

Objective: This article provides the test–retest reliability and Reliable Change Indices (RCIs) of the Philips IntelliSpace Cognition (ISC) platform, which contains digitized versions of well-established neuropsychological tests.

Method: 147 participants (ages 19 to 88) completed a digital cognitive test battery on the ISC platform or paper-pencil versions of the same test battery during two separate visits. Intraclass correlation coefficients (ICC) were calculated separately for the ISC and analog test versions to compare reliabilities between administration modalities. RCIs were calculated for the digital tests using the practice-adjusted RCI and standardized regression-based (SRB) method.

Results: Test–retest reliabilities for the ISC tests ranged from moderate to excellent and were comparable to the test–retest reliabilities for the paper-pencil tests. Baseline test performance, retest interval, age, and education predicted test performance at visit 2 with baseline test performance being the strongest predictor for all outcome measures. For most outcome measures, both methods for the calculation of RCIs show agreement on whether or not a reliable change was observed.

Conclusions: RCIs for the digital tests enable clinicians to determine whether a measured change between assessments is due to real improvement or decline. Together, this contributes to the growing evidence for the clinical utility of the ISC platform.

ARTICLE HISTORY

Received 19 September 2023

Accepted 11 January 2024

Published online 15 February 2024

KEYWORDS

Neuropsychology; cognitive tests; digital technology; test–retest reliability; intraclass correlation; reliable change index; practice effects

Introduction

Recent years have seen a steady rise of digital cognitive test batteries that aim to capitalize on the enormous advantages of technology-enabled digital assessments compared to traditional paper-pencil neuropsychological assessments (Arrioux et al., 2017; Bauer et al., 2012; Chan et al., 2021; Feenstra et al., 2017; Germine et al., 2019; Kessels, 2019; Riordan et al., 2013; Zygouris & Tsolaki, 2015). Despite the existence of

CONTACT Laura Klaming  laura@klaming.com  Digital Cognitive Diagnostics, Philips Healthcare, High Tech Campus 34, 5656 AE Eindhoven, The Netherlands.

© 2024 Informa UK Limited, trading as Taylor & Francis Group

numerous digital cognitive test batteries and the ubiquity of digital technology in everyday life, the adoption of digital cognitive tests in neuropsychological practice initially progressed slowly (Miller & Barr, 2017; Rabin et al., 2014) but will likely increase rapidly given recent validation efforts of various digital cognitive test batteries (Cole et al., 2013; Gualtieri & Johnson, 2006; Littleton et al., 2015; Nakayama et al., 2014; Staffaroni et al., 2020; Weintraub et al., 2013), which has been accelerated further by an increasing need for remote healthcare.

One of the crucial prerequisites for the adoption of digitized cognitive testing in clinical practice is the establishment of psychometric data for digital cognitive tests (Rabin et al., 2014; Schmand, 2019). Validity and test–retest reliability of digital cognitive tests are of critical importance to determine their appropriateness for clinical use. This is of particular importance given that digital cognitive tests may have important differences from their paper-pencil counterparts, e.g. in terms of response format, that may result in measuring functions that are not equivalent (O'Brien et al., 2022). We have already provided evidence of the validity of Philips IntelliSpace Cognition (ISC), a digital cognitive assessment tool consisting of digitized versions of existing paper-pencil cognitive tests (Vermeent et al., 2022). Here, we will extend these findings by presenting the test–retest reliability of ISC. Test–retest reliability is the extent to which a test produces consistent results across two or more administrations to the same person. Test–retest reliability is one of the fundamental psychometric properties that need to be established for digital cognitive assessment tools such as ISC. It is crucial in a clinical context, because higher test–retest reliability provides the clinician with greater confidence that a change in score is due to the cognitive function measured rather than instability of the cognitive test.

ISC is a digital cognitive assessment platform classified as a Food and Drug Administration (FDA, 2020) Class II medical device (FDA Medical Devices, 21 C.F.R. 55 882.1470). The ISC battery consists of digitized versions of commonly used and well-established paper-pencil cognitive tests (Rabin et al., 2016). In addition, ISC features a digital adaptation of the Mini Mental State Examination 2nd Edition (MMSE-2; Folstein et al., 2010) as well as digitized versions of the Patient Health Questionnaire (PHQ-9) and the General Anxiety Disorder-7 (GAD-7).

Here, we investigate the test–retest reliability of the digital cognitive tests on ISC and compare it to the test–retest reliability of the same paper-pencil cognitive tests to evaluate the clinical viability of ISC. Direct comparisons between digital and analog test–retest coefficients are scarce in the literature. Instead, empirically obtained digital test–retest coefficients for digital cognitive tests are typically compared to analog coefficients reported in the literature (e.g., Gualtieri & Johnson, 2006; Littleton et al., 2015; Nakayama et al., 2014). Such comparisons are difficult to interpret because of differences in samples, test versions, and retest intervals. We mitigated these limitations by using a between-subjects design with two groups, one receiving the digital cognitive tests on two occasions and one receiving the paper-pencil versions of the same cognitive tests on two occasions. We report test–retest reliabilities of the main outcome measures of each cognitive test on ISC and compare them to the test–retest reliability coefficients of the analog versions of the same cognitive tests. A secondary purpose of this paper is to provide reliable change indices (RCIs) for the cognitive tests on ISC. RCIs provide valuable information to clinicians by giving estimates of the probability

that a given difference in score between two or more assessments is due to a change in the patient's ability rather than measurement error. This can be important when tracking the progression of neurological diseases such as Alzheimer's disease (AD) and related neurodegenerative syndromes, as well as when monitoring the impact of certain treatments or interventions intended to improve cognitive function.

Methods

Participants

The two studies described in this article were part of two larger studies with a broader scope including a total of 922 participants (see also Vermeent et al., 2022). Participants were recruited through a research recruitment agency, following a sampling plan that stratified for age group, gender, education level and racial/ethnic background according to the United States demographic data (US Census Bureau, 2017). For the purpose of this article, we only report on the results of participants who received two assessments with the same modality, either both assessments on ISC ($N=97$) or both assessments on paper ($N=49$).

Figure 1 shows a flowchart for both studies. Individuals who were eligible to participate in the study were randomly assigned to a group with minimization for demographic characteristics. In study 1, 53 participants were assigned to the ISC group and performed a digital cognitive test battery on the ISC platform during both of their visits, while 49 participants were assigned to the paper-pencil group and performed analog tests during both of their visits. In study 2, 45 participants were assigned to the ISC group and performed a digital cognitive test battery on the ISC platform. This resulted in an ISC group of 98 participants and an analog group of 49 participants. One participant was removed from the ISC group because his retest interval was more than three SD longer than the mean retest interval, and it may be assumed that because of the long retest interval his performance on the cognitive tests is not comparable to the rest of the group. This resulted in an ISC group of 97 participants.

The total sample for this study ($N=146$) consisted of volunteers with no clinical diagnosis of mild cognitive impairment or dementia and no self-reported cognitive complaints (mean MMSE-2 score = 27.7, $SD=1.8$, range = 22–30) between the ages of 19 and 88 years old who were living independently at the time of data collection. Their primary language was English, and they had normal or corrected-to-normal eyesight and hearing on the days of testing, as well as normal fine and gross motor ability.

The exclusion criteria were: (1) self-reported presence of neurological, autoimmune, psychotic, mood, anxiety, autism spectrum, substance use, or learning disorder or intellectual disability, (2) recent functional changes in reference to activities of daily living (ADLs), (3) aphasia or any cognitive difficulties for which participants were seeking medical help at the time of the study, (4) admission to a hospital, or residence in an assisted living facility, nursing home or psychiatric facility at the time of the study, (5) history of head trauma with loss of consciousness greater than 20 min, (6) history of a medical event requiring resuscitation, history of electroconvulsive therapy or radiation to the central nervous system, (7) a neuropsychological assessment in

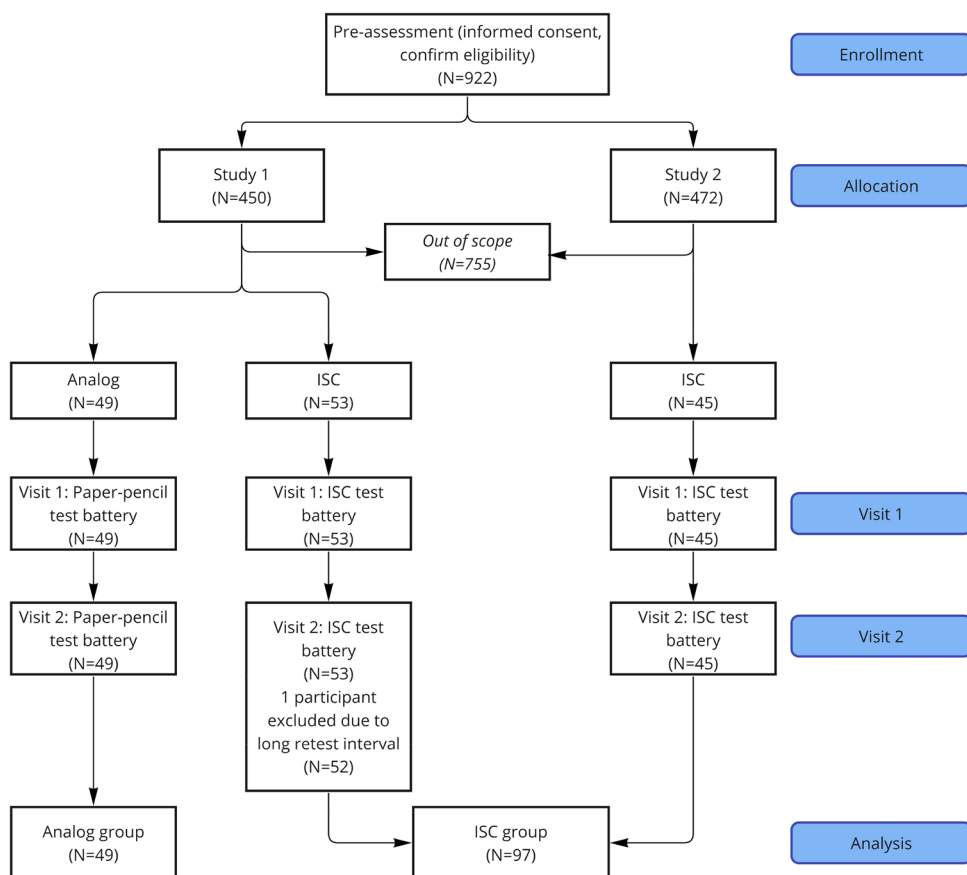


Figure 1. Study flow chart. Note. ISC: IntelliSpace Cognition.

the six months preceding the study, (8) use of medication that might impact test performance (e.g., anticonvulsants, benzodiazepines) or chemotherapy treatment in the 2 months preceding the study. Recruitment and testing took place at four locations in the US: Totowa (NJ), Orlando (FL), Philadelphia (PA), and Sacramento (CA).

Measures

An overview of the outcome measures per test, and a description of the administration and scoring differences between the two administration modalities is presented in Table 1. The tests were administered in the following fixed order for each participant at each visit: Mini Mental Status Examination (2nd edition, MMSE-2), Rey Auditory Verbal Learning Test (RAVLT) Learning Trials, RAVLT Interference, RAVLT Immediate Recall, Trail Making Test (TMT) A, TMT B, Clock Test (CT) Drawing, CT Copy, Star Cancellation Test (SCT), RAVLT Delayed Recall, RAVLT Recognition, Rey-Osterrieth Complex Figure Test (ROCFT) Copy, Letter Fluency Test (LFT, also known as Controlled Oral Word Association Test), ROCFT Immediate Recall, Digit Span Test (DST) Forward, DST Backward, and Category Fluency Test (CFT). The test order was fixed for all

Table 1. Overview of tests and outcome measures.

Test	Outcome measures	Administration	Scoring
MMSE-2	Total score (0–30)	IAW Folstein et al. (2010). Instructions adapted to reflect digital format for ISC and provided orally by test administrator.	ISC: Immediate scoring in-tool by test administrator. Analog: Transcription by test administrator; retrospective scoring by neuropsychologist.
Clock Test Copy Clock Test Drawing	ISC: Total score (0–22) Analog: Total score (0–10)	IAW Strauss et al. (2006). Instructions adapted to reflect digital format for ISC and provided by standardized built-in audio.	ISC: Scoring by designated algorithm. Analog: Retrospective scoring by neuropsychologist according to Manos and Wu (1994).
RAVLT Learning Trials	Total score (0–75)	IAW Schmidt (1996). Instructions adapted to reflect digital format for ISC and provided orally by test administrator. ISC: Response captured by audio recording.	ISC: Scoring by designated algorithm.
RAVLT Interference	Total score (0–15)		Analog: Retrospective scoring by neuropsychologist.
RAVLT Immediate Recall			
RAVLT Delayed Recall			
RAVLT Recognition	Total score (0–30)	Analog: Response captured by test administrator transcription.	
TMT A TMT B	Duration (in seconds)	IAW Strauss et al. (2006). Instructions adapted to reflect digital format for ISC and provided by standardized built-in audio.	ISC: Scoring by designated algorithm. Analog: Immediate timing by test administrator (using stopwatch).
Star Cancellation Test	Total score (0–54), Duration (in seconds) Duration (in seconds) per correct cancellation	IAW Wilson et al. (1987). Instructions adapted to reflect digital format for ISC and provided orally by test administrator.	ISC: Scoring by designated algorithm. Analog: Immediate timing by test administrator (using stopwatch); retrospective scoring by neuropsychologist.
ROCFT Copy ROCFT Immediate Recall	Total score (0–36)	Instructions adapted to reflect digital format for ISC and provided by standardized built-in audio.	ISC: Scoring by designated algorithm. Analog: Retrospective scoring by neuropsychologist according to Meyers & Meyers (1995).
Letter Fluency Test	Total score trials 1–3	IAW Benton & Hamsher (1989), Schmand et al. (2008). Instructions adapted to reflect digital format for ISC and provided orally by test administrator. ISC: Response captured by audio recording. Analog: Response captured by test administrator transcription.	ISC: Scoring by designated algorithm. Analog: Transcription by test administrator; retrospective scoring by neuropsychologist.

(Continued)

Table 1. Continued.

Test	Outcome measures	Administration	Scoring
Category Fluency Test Animals	Total score per trial, Sum of scores trials 1–3	IAW Benton & Hamsher (1989). Instructions adapted to reflect digital format for ISC and provided orally by test administrator. ISC: Response captured by audio recording. Analog: Response captured by test administrator transcription.	ISC: Scoring by designated algorithm.
Category Fluency Test Vegetables			Analog: Transcription by test administrator; retrospective scoring by neuropsychologist.
Category Fluency Test Fruit & Furniture			
Digit Span Forward Digit Span Backward	Total score (0–12)	IAW Lezak et al. (2004). Instructions adapted to reflect digital format for ISC and provided by standardized built-in audio. ISC: Item presented visually; response captured by participant keypad presses. Analog: Item presented orally; response captured by test administrator transcription of verbal response.	ISC: Scoring by designated algorithm. Analog: Transcription by test administrator; retrospective scoring by neuropsychologist.

Note. ISC = IntelliSpace Cognition; MMSE-2 = Mini-Mental State Examination 2nd Edition; RAVLT = Rey Auditory Verbal Learning Test; TMT = Trail Making Test; ROCFT = Rey-Osterrieth Complex Figure Test.

participants for three main reasons. First, a fixed test order ensured that test interval requirements were adhered to for the RAVLT and ROCFT recall trials (e.g., comparable time intervals between the learning and (delayed) recall trials). Second, the order in which tests were administered ensured that there were no verbal tests between the different RAVLT trials and no drawing tests between the ROCFT trials to minimize interference of test outcomes on the RAVLT and ROCFT due to other tests. And third, a fixed test order ensured that test order did not influence test–retest reliability. The analog versions of the tests were administered in the same fixed order as the ISC tests, according to standardized administration rules as used in clinical practice.

The digital cognitive tests were administered via the Philips IntelliSpace Cognition (ISC) platform using an Apple iPad Pro tablet (3rd generation) running on iOS 12 with a screen size of 12.9 inches and resolution of 2732 × 2048. Drawings were made using an Apple Pencil (2nd generation) for the iPad Pro. For tests that required a verbal response, audio recordings were captured using the iPad's internal microphone. All tests, except the MMSE-2, were automatically scored by machine learning algorithms that have been validated and described elsewhere (see Vermeent et al., 2022).

Procedure

Ethics approval for the studies was obtained from the Internal Committee of Biomedical Experiments of Philips as well as from an external institutional review board (Western

IRB/WCG IRB). The studies were conducted according to the principles of the Declaration of Helsinki and per the requirements of the ISO 14155 standard (Good Clinical Practice). The studies are registered on ClinicalTrials.gov (NCT03801382 and NCT04729257).

For each participants, the total duration per visit was approximately 1.5 h, and the visits were between 13 and 56 days apart ($M = 18$ days, $SD = 7$ days, median = 15 days). At the beginning of the visit, participants signed an informed consent (only at visit 1), the test administrator confirmed eligibility of the participant by verifying all inclusion and exclusion criteria and completed either the ISC or analog version of the test battery. During the first visit, some of the participants also filled in digital familiarity and usability questionnaires, but the outcomes of these questionnaires are not discussed in this article.

All assessments took place in a dedicated examination room at one of the four study locations. Sources of light reflections on the iPad (e.g., from a window or a lamp) were eliminated. All tests were administered by a licensed psychologist with clinical experience in neuropsychological testing. For each individual participant, both assessments were administered by the same test administrator. Participants entered the study voluntarily and received 75 US dollar compensation for each visit. Participants did not receive feedback about their performance on the tests.

After data collection was completed, a total of 21 ISC test scores (<1% of all test scores for the ISC group) were excluded case wise due to technical difficulties with recording the responses. For the Category Fluency Test, four scores were missing for each of the categories and the total score, resulting in 16 test scores that were excluded for this test. For the Letter Fluency Test, two total scores were missing. For the RAVLT, one Delayed Recall score was missing. For the TMT A and B, one score was missing for each. A total of 192 analog scores (9% of all test scores for the analog group) were excluded case wise due to missing values. For the Category Fluency Test, 24 scores were missing for each of the categories and the total score, resulting in 192 test scores (49%) that were excluded for this test. The reason for the high number of missing values for the Category Fluency Test was logistical; due to resource constraints, it was decided to only score half of the sample, and hence data were only available for half of the sample included in this study.

Analyses

All analyses were conducted using Python 3.11.0 (including libraries: Pandas 1.5.1, Numpy 1.23.4, Pingouin 0.5.2, SciPy 1.9.3, Statsmodels 0.13.5). Independent samples *t*-tests, Mann-Whitney *U* tests, or chi-square tests were conducted to compare the two groups on the demographic variables. For each outcome measure, intraclass correlation coefficients (ICC) for absolute agreement with two-way mixed effects (Koo & Li, 2017) were calculated with the Pingouin package (0.5.2). ICC rather than Pearson correlation coefficients were chosen because they account for individual variability. The strength of test–retest reliability was assessed following the guidelines set out by Koo and Li (2017): <.50 indicating poor reliability, between .50 and .75 indicating moderate reliability, between .75 and .90 indicating good reliability, and >.90 indicating excellent reliability. In addition, we report the raw score differences between visit 1

and visit 2 for each outcome measure for both groups. ANCOVAs including age as covariate were conducted to analyze between group differences in difference scores between visit 1 and visit 2. For these comparisons, alpha was Bonferroni corrected (.05/21 = .002). For the ISC group, the scoring of the Clock Test was based on an algorithm and scores ranged from 0 to 22, while for the analog group, the scoring of the Clock Test was performed manually according to Manos and Wu (1994) with scores ranging from 0 to 10. For the ANCOVAs, scores were therefore normalized.

For the ISC group, mean practice effects were calculated by subtracting the group mean of visit 1 from the group mean of visit 2. Paired-samples *t*-tests were conducted to analyze significant practice effects for each outcome measure.

For the ISC group, reliable change indices (RCIs) were calculated for all cognitive tests from visit 1 to visit 2. RCI which has originally been proposed by Jacobson and Truax (1991) is a methodology used to determine if an observed change between visits is beyond what would be expected by chance. Different methods to calculate RCI have been developed (see e.g. Hinton-Bayre, 2010). We used two methods for RCI, namely RCI by Chelune and colleagues (1993) correcting for practice effects with Iverson's standard error of the difference,¹ and the standardized regression-based (SRB) method developed by McSweeney and colleagues (1993). SRB uses linear regression to predict retest (visit 2) scores for individuals based on their baseline (visit 1) scores and other potential predictors. A change score is calculated from an individual's predicted and observed score at visit 2 and divided by the standard error of the estimate of the regression model. Age, years of education, retest interval, and visit 1 score were included as independent variables. Multiple linear regression analyses were conducted with these predictors. Outcome measures that have been measured in seconds were log-transformed to improve normality. 90% and 95% confidence intervals are reported for the RCI_{Chelune} method. These were obtained by multiplying the standard error of the difference by 1.64 (90%) or 1.96 (95%) and adding the practice effect.

Results

The sample consisted of a total of 146 volunteers with a mean age of 61 years ($SD=13$) and a mean of 14 years of education ($SD=3$). The sample included 78 women and 68 men with the majority of participants being White ($N=103$), followed by Hispanic ($N=17$), Black ($N=15$), and other ($N=11$).

As can be seen in Table 2, there were no significant differences between the ISC and analog groups in gender ($\chi^2(1) = .01, p=.9$), ethnicity ($\chi^2(3) = .08, p=.99$), years of education ($z=-1.6, p=.11$), study location ($\chi^2(3) = 1.99, p=.57$), MMSE-2 ($t(144) = 1.32, p=.19$), and time interval between visit 1 and visit 2 ($z=-0.7, p=.48$). Participants in the ISC group were significantly younger than participants in the analog group ($z=-2.2, p<.05$).

Table 3 presents an overview of the test-retest reliability coefficients for the two groups and for each of the outcome measures included in the study. For the ISC group, of the 21 outcome measures, 9 had moderate test-retest reliability, 11 had good test-retest reliability, and 1 had excellent test-retest reliability. For the analog group, of the 21 outcome measures, 3 had poor test-retest reliability, 5 had moderate

Table 2. Demographics, MMSE-2, study location, and retest interval.

	ISC group (N=97)	Analog group (N=49)	Test statistics	p
Age, mean (SD)	59 (15)	66 (7)	$z = -2.2$	<.05
Years of education, mean (SD)	14.6 (2.7)	13.8 (2.2)	$z = -1.6$.11
Gender, % female	53	55	$\chi^2(1) = .01$.91
Ethnicity			$\chi^2(3) = .08$.99
White (%)	71.1	69.4		
Hispanic (%)	11.3	12.2		
Black (%)	10.3	10.2		
Other (%)	7.2	8.2		
MMSE-2 score at visit 1, mean (SD)	27.5 (1.9)	27.9 (1.6)	$t(144) = 1.32$.19
Study location			$\chi^2(3) = 1.99$.57
Totowa, NJ (%)	20.6	30.6		
Orlando, FL (%)	24.7	22.4		
Philadelphia, PA (%)	27.8	26.5		
Sacramento, CA (%)	26.8	20.4		
Test-retest interval in days, mean (SD)	18 (6)	19 (9)	$z = -0.7$.48

Note. ISC = IntelliSpace Cognition; MMSE-2 = Mini Mental State Examination 2nd edition.

test-retest reliability, 12 had good test-retest reliability, and 1 had excellent test-retest reliability.

We also compared the mean differences between visit 1 and 2 between the two groups. There were no significant differences between the ISC and analog group for any of the outcome measures with a Bonferroni corrected alpha (.002).

RCIs were calculated between visit 1 and visit 2 for the ISC group. Mean test-retest intervals were 18 days (SD = 6 days). Raw scores at visit 1 and visit 2, test-retest reliability (ICCs), SE_{Diff} mean group practice effects, and 90% and 95% confidence intervals for the $RCI_{Chelune}$ method are provided in Table 4. RCIs were rounded to the nearest whole number, except for outcomes measured in seconds. As can be seen in Table 4, some cognitive tests have very small group mean practice effects. For 13 of the 21 outcome measures, there was a statistically significant difference between visit 1 and visit 2.

Using linear regression, SRB normative trajectories were computed with age, years of education, retest interval, and visit 1 performance as independent variables. Table 5 provides an overview of the SRB equations predicting visit 2 performance. For all outcome measures, visit 1 performance was the strongest predictor of visit 2 performance. Age was found to be a significant predictor in the regression models for the RAVLT total score, RAVLT Interference, RAVLT Immediate Recall, RAVLT Delayed Recall, TMT A and B, and the Letter Fluency Test. Education was a significant predictor in the regression model for the TMT B. Retest interval was a significant predictor in the regression models for the Star Cancellation duration, Star Cancellation duration per correct cancellation, the Category Fluency Test Vegetables, and the Digit Span forward.

To illustrate the use of SRB change scores in the interpretation of cognitive change, Table 6 provides visit 1, visit 2, difference scores, and SRB change scores for a 75-year-old participant with 14 years of education and a retest interval was 14 days. An example of an SRB change score calculation² is provided for illustration: the participant had a score of 8 on the RAVLT Immediate Recall on visit 1 and a score of 11 on visit 2. The predicted visit 2 score is $7.65 + (-0.05 \times 75) + (-0.1 \times 14) + (.04 \times 14) + (.75 \times 8) = 9.06$. The SRB change score is $(11 - 9.06) / 2.07 = .94$. This change

Table 3. Test–retest statistics of each test for ISC and analog group and between group differences in visit 2 – visit 1.

Test	Outcome measure	ISC group (N=97) ¹		Analog group (N=49) ¹		Between group difference in visit 2 – visit 1		
		Mean (SD)	ICC [95% CI]	Mean (SD)	ICC [95% CI]	F	p	Eta squared
CT ²	Copy							
	Visit 1	19.2 (2.2)	.52 ^b [.29, .68]	9.8 (.5)	.12 ^a [-0.56, .5]	F(1, 143)=4.88 =1.52	.03	.03
Visit 2	19.9 (1.8)		9.7 (.8)					
	Drawing					F(1, 143)	.22	.01
	Visit 1	17.9 (3.2)	.64 ^b [.43, .77]	8.8 (2)	.38 ^a [-0.11, .65]			
	Visit 2	19.2 (2.9)		9 (1.2)				
	Learning Trials total score					F(1, 143)=.85	.36	.01
Visit 1	42.2 (9)	.79 ^c [-0.02, .92]	41.8 (7.8)	.76 ^c [.15, .91]				
	Visit 2	49.5 (11.1)		47.6 (9.3)				
	Interference					F(1, 143)=.82	.37	.01
Visit 1	5.3 (1.9)	.69 ^b [.53, .79]	4.8 (1.5)	.70 ^b [.48, .83]				
	Visit 2	5.4 (2.1)		4.6 (1.5)				
	Immediate Recall					F(1, 143)=.86	.36	.01
Visit 1	7.7 (3)	.77 ^c [.2, .9]	7.9 (3.0)	.79 ^c [.46, .9]				
	Visit 2	9.7 (3.2)		9.3 (3.0)				
	Delayed recall					F(1, 142)=.01	.94	.00
Visit 1	7.8 (3.4)	.67 ^b [.39, .81]	7.6 (3.2)	.88 ^c [.21, .96]				
	Visit 2	9.7 (3.8)		9.3 (3.1)				
	Recognition					F(1, 143)=.71	.40	.01
Visit 1	27.2 (2.4)	.73 ^b [.57, .83]	27.1 (2.6)	.83 ^c [.7, .91]				
	Visit 2	27.9 (2.2)		27.7 (2.1)				
	A duration (seconds)					F(1, 142)=2.06	.15	.01
Visit 1	40.2 (15.8)	.71 ^b [.56, .81]	31.1 (9.8)	.56 ^b [.22, .75]				
	Visit 2	36 (12.4)		30.2 (8.4)				
	B duration (seconds)					F(1, 142)=.23	.63	.00
Visit 1	103.5 (59.3)	.68 ^b [.51, .78]	85.0 (67.8)	.81 ^c [.66, .89]				
	Visit 2	90.3 (45.6)		75.6 (41.2)				
	Total score					F(1, 143)=1.27	.26	.01
Visit 1	53 (1.6)	.60 ^b [.41, .74]	53.3 (0.9)	.39 ^a [-0.07, .66]				
	Visit 2	53.1 (1.4)		53.2 (1.2)				
	Duration (seconds)					F(1, 143)=3.44	.07	.02
Visit 1	45.7 (14)	.83 ^c [.73, .89]	42.4 (12.5)	.88 ^c [.79, .93]				
	Visit 2	42.3 (12.5)		41.7 (11.7)				
	Duration (seconds) per correct cancellation					F(1, 143)=4.25	.04	.03
Visit 1	.9 (3)	.82 ^c [.72, .89]	.8 (.2)	.89 ^c [.8, .94]				
	Visit 2	.8 (2)		.8 (2)				
	Copy					F(1, 143)=.26	.61	.00
Visit 1	30 (6.3)	.74 ^b [.61, .82]	26.2 (6.9)	.81 ^c [.67, .89]				
	Visit 2	29.4 (7.2)		26.5 (5.8)				
	Immediate recall					F(1, 143)=1.7	.19	.01
Visit 1	11.1 (6.7)	.79 ^c [.63, .88]	11.9 (6.8)	.92 ^d [.83, .96]				
	Visit 2	13.8 (7.8)		13.4 (6.5)				
	Total score trials 1-3					F(1, 141)=2.56	.11	.02
Visit 1	38.7 (10.1)	.83 ^c [.69, .9]	39.2 (12.1)	.89 ^c [.81, .94]				
	Visit 2	42 (9.8)		39.7 (12.8)				
	Animals					F(1, 115)=.16	.69	.00
Visit 1	19.9 (6)	.88 ^c [.81, .92]	20.6 (5.8)	.68 ^b [.28, .86]				
	Visit 2	20 (5.9)		20.2 (4.7)				
	Vegetables					F(1, 115)=.47	.49	.00
Visit 1	12.8 (3.9)	.80 ^c [.71, .87]	14.1 (3.3)	.69 ^b [.31, .87]				
	Visit 2	13.3 (4.3)		13.7 (3.9)				
	Fruit & Furniture					F(1, 115)=2.76	.09	.02
Visit 1	12.8 (3)	.77 ^c [.66, .85]	12.5 (2.9)	.70 ^b [.33, .87]				
	Visit 2	12.5 (3)		13.0 (2.8)				
	Total score trials 1-3					F(1, 115)=.00	.96	.00
Visit 1	45.5 (10.6)	.93 ^d [.89, .95]	47.2 (10.0)	.79 ^c [.53, .91]				
	Visit 2	45.8 (10.8)		47.0 (8.1)				

(Continued)

Table 3. Continued.

Test	Outcome measure	ISC group (N=97) ¹		Analog group (N=49) ¹		Between group difference in visit 2 – visit 1		
		Mean (SD)	ICC [95% CI]	Mean (SD)	ICC [95% CI]	F	p	Eta squared
DST	Forward							
	Visit 1	7.5 (2.2)	.77 ^c [.66, .85]	8.7 (2.1)	.86 ^c [.75, .92]	F(1, 143)=2.12	.15	.02
	Visit 2	8 (2.2)		8.6 (1.9)				
	Backward					F(1, 143)=.14	.71	.00
Visit 1	6.9 (2.5)	.76 ^c [.64, .84]	6.5 (1.9)	.80 ^c [.65, .89]				
Visit 2	6.9 (2.6)		6.6 (2.2)					

Note. ISC = IntelliSpace Cognition; RAVLT = Rey Auditory Verbal Learning Test; TMT = Trail Making Test; ROCFT = Rey-Osterrieth Complex Figure Test; Letter Fluency Test = Controlled Oral Word Association Test.

¹For the ISC group, 16 scores were missing for the CFT, two for the LFT, one for the RAVLT Delayed Recall, one for the TMT A, and one for the TMT B. For the analog group, a total of 192 scores were missing for the CFT.

²For ISC group, scoring based on algorithm (0–22), for analog group, scoring according to Manos and Wu (1994; range 0–10). For ANCOVAs, scores were normalized.

^aPoor; ^bModerate; ^cGood; ^dExcellent (Koo & Li, 2017).

Table 4. Visit 1 and 2 scores, test-retest reliability (ICC), SEDiff, practice effects, and 90% and 95% confidence intervals for the RCiChelune method for the ISC outcome measures.

Test		Visit 1 mean (SD)	Visit 2 mean (SD)	ICC	SE _{Diff}	Practice effect	RCI confidence interval			
							90%		95%	
							det	imp	det	imp
CT	Copy	19.2 (2.2)	19.9 (1.8)	.52	1.95	.6*	-3	+4	-3	+5
	Drawing	17.9 (3.2)	19.2 (2.9)	.64	2.60	1.3*	-3	+6	-4	+6
RAVLT	Learning trials total score	42.2 (9)	49.5 (11.1)	.79	6.59	7.4*	-4	+18	-6	+20
	Interference									
	Immediate recall	5.3 (1.9)	5.4 (2.1)	.69	1.59	.1	-3	+3	-3	+3
	Delayed recall	7.7 (3)	9.7 (3.2)	.77	2.12	2.0*	-2	+6	-2	+6
TMT	Recognition	7.8 (3.4)	9.7 (3.8)	.67	2.93	1.9*	-3	+7	-4	+8
	A duration (seconds)	27.2 (2.4)	27.9 (2.2)	.73	1.70	.8*	-2	+4	-3	+4
	B duration (seconds)	40.2 (15.8)	36 (12.4)	.71	10.76	-4.2*	+13.5	-21.9	+16.9	-25.3
SCT	Total score	103.5 (59.3)	90.7 (45.5)	.68	42.51	-13.2*	+57.1	-82.8	+70.5	-96.2
	Duration (seconds)	53 (1.6)	53.1 (1.4)	.60	1.35	.1	-2	+2	-3	+3
	Duration (seconds) per correct cancellation	45.7 (14.1)	42.3 (12.5)	.83	7.80	-3.4*	+9.4	-16.3	+11.9	-18.7
ROCFT	Copy	.9 (3)	.8 (2)	.82	.15	-0.1*	+2	-0.3	+2	-0.4
	Immediate recall	30 (6.3)	29.4 (7.2)	.74	4.90	-0.5	-9	+8	-10	+9
LFT	Copy	11.1 (6.7)	13.8 (7.8)	.79	4.68	2.7*	-5	+10	-7	+12
	Total score trials 1–3	38.7 (10.1)	42 (9.8)	.83	5.86	3.3*	-6	+13	-8	+15
CFT	Animals	19.9 (6)	20 (5.9)	.88	2.98	.1	-5	+5	-6	+6
	Vegetables	12.8 (3.9)	13.3 (4.3)	.80	2.59	.5	-4	+5	-5	+6
	Fruit and furniture	12.8 (3)	12.5 (3)	.77	2.03	-0.3	-4	+3	-4	+4
	Total score trials 1–3	45.5 (10.6)	45.8 (10.8)	.93	4.07	.3	-6	+7	-8	+8
DST	Forward	7.5 (2.2)	8 (2.2)	.77	1.47	.5*	-2	+3	-2	+3
	Backward	6.9 (2.5)	6.9 (2.6)	.76	1.76	.0	-3	+3	-4	+4

Note. det = deteriorated, imp = improved. Statistically significant at .05.

Table 5. Regression models and coefficients for the ISC outcome measures.

Test		Adj R ²	SEE	Constant	Age	Education	Retest interval	Visit 1
CT	Copy	.13	1.50	14.14	-0.01	-0.01	.05	.28*
	Drawing	.24	2.21	9.81	.01	.04	.00	.44*
RAVLT	Learning Trials total score	.71	6.85	18.03	-0.11*	-0.21	-0.03	.98*
	Interference	.29	1.53	5.13	-0.03*	-0.08	.01	.54*
	Immediate recall	.62	2.07	7.65	-0.05*	-0.10	.04	.75*
	Delayed recall	.34	2.82	6.93	-0.05*	.04	.01	.61*
	Recognition	.35	1.49	12.64	.00	.00	.00	.56*
TMT	A duration (seconds)	.55	8.71	.66	.01*	-0.01	.00	log10(.39)*
	B duration (seconds)	.49	33.52	.99	.01*	-0.01*	.00	log10(.44)*
SCT	Total score	.19	1.14	36.38	-0.02	.06	-0.01	.32*
	Duration (seconds)	.54	6.99	.39	.00	.00	.01*	log10(.69)*
	Duration (seconds) per correct cancellation	.54	.13	-0.14	.00	.00	.01*	log10(.69)*
ROCF1	Copy	.33	4.89	7.19	-0.02	.05	.12	.67*
	Immediate recall	.49	4.77	7.87	-0.05	.11	-0.08	.79*
LFT	Total score trials 1-3	.57	5.54	18.08	-0.12*	.26	.08	.67*
CFT	Animals	.61	2.86	3.02	-0.03	.27	.05	.70*
	Vegetables	.49	2.58	3.49	-0.03	-0.05	.15*	.74*
	Fruit and furniture	.42	1.89	7.36	-0.03	.00	-0.08	.64*
	Total score trials 1-3	.75	4.02	8.53	-0.07	.03	.11	.85*
DST	Forward	.44	1.36	3.72	-0.02	-0.01	.06*	.59*
	Backward	.39	1.68	1.44	-0.02	.15	.04	.56*

Note. SEE = standard error of the estimate of the regression model. For outcome measures that have been log transformed, regression equations need to be back-transformed (i.e. $10^{\wedge}\text{regression equation}$). Significant predictor.

score falls within ± 1.645 (90% confidence interval) and therefore indicates that the participant was stable from visit 1 to visit 2. Similarly, when looking at Table 4 which provides the RCI for the RCI_{Chelune} method, a difference score of +3 on the RAVLT Immediate Recall falls within the 90% confidence interval ranging from -2 to +6 indicating that the participant was stable.

The last two columns in Table 6 show whether the participant had a reliable change based on the SRB method and the RCI_{Chelune} method using a 90% confidence interval. For most outcome measures, both methods show agreement on whether there was a reliable change or not. However, for four outcome measures, the two methods provide different indications as to whether the observed change is truly an improvement or deterioration. Rather than indicating that one of the two methods is superior, this example highlights the complexities involved in identifying reliable cognitive change.

Discussion

The test-retest reliabilities of the cognitive tests on ISC ranged from moderate to excellent and were found to be comparable to the test-retest reliabilities of the same

Table 6. Case example's visit 1, visit 2, difference scores, and SRB change scores for ISC outcome measures. The last two columns indicate whether the participants had a reliable change based on the SRB method and the RCI_{Chelune} method using a 90% confidence interval (± 1.645).

Test		Visit 1	Visit 2	Difference	SRB change score	SRB reliable change	RCI_{Chelune} reliable change
CT	Copy	22	19	-3	-0.74	No change	Deteriorated
	Drawing	20	20	0	.02	No change	No change
RAVLT	Learning trials	43	52	9	.50	No change	No change
	total score						
	Interference	5	3	-2	-1.05	No change	No change
	Immediate recall	8	11	3	.94	No change	No change
TMT	Delayed recall	8	10	2	.44	No change	No change
	Recognition	29	29	0	-0.06	No change	No change
	A duration (seconds)	30.3	41.4	11.1	.78	No change	No change
	B duration (seconds)	88.7	73.1	-15.5	-1.02	No change	No change
SCT	Total score	49	53	4	1.53	No change	Improved
	Duration (seconds)	47.4	33.6	-13.8	-1.31	No change	No change
	Duration (seconds) per correct cancellation	.97	.63	-0.34	-1.64	Improved	Improved
ROCF	Copy	31	32	1	.65	No change	No change
	Immediate recall	6	21	15	2.43	Improved	Improved
LFT	Total score trials 1-3	45	43	-2	-0.18	No change	No change
CFT	Animals	16	12	-4	-1.56	No change	No change
	Vegetables	14	10	-4	-1.16	No change	Deteriorated
	Fruit and furniture	13	14	1	.89	No change	No change
	Total score trials 1-3	43	36	-7	-1.44	No change	Deteriorated
DST	Forward	8	8	0	.21	No change	No change
	Backward	3	5	2	.43	No change	No change

paper-pencil cognitive tests. Moreover, the test-retest reliabilities found for the digital cognitive tests on ISC are in line with the test-retest reliabilities found in previous research with paper-pencil cognitive tests (Calamia et al., 2013; Duff, 2014; Hammers et al., 2021, 2022).

As described earlier, rather than directly comparing digital and analog test-retest coefficients, studies typically compare empirically obtained digital test-retest reliabilities for digital cognitive tests to analog coefficients reported in the literature (e.g., Gualtieri & Johnson, 2006; Littleton et al., 2015; Nakayama et al., 2014). Such comparisons are difficult to interpret because of differences in samples, test versions, and retest intervals. In the present study, we mitigated these limitations by using a between-subjects design with comparable groups in terms of gender, education level, ethnicity, MMSE-2 score, and retest interval. We found excellent test-retest reliabilities for one outcome measure in the ISC group and for one outcome measure in the analog group, good test-retest reliabilities for 11 outcome measures in the ISC group and for 12 outcome measures in the analog group, moderate test-retests reliabilities for 9 outcome measures in the ISC group and for 5 outcome measures in the analog group. We did not find poor test-retest reliabilities for any outcome measure in the ISC group, compared to three poor test-retest reliabilities in the analog group. It seems unlikely that the differences found in test-retest reliabilities

are due to differences in administration modality since differences were found for tests that require direct interaction with the iPad (e.g., drawing tests such as the Clock Test) as well as for tests that require minimal interaction with the iPad (e.g., verbal tests such as the Category Fluency Test and RAVLT which automatically record verbal responses). Moreover, differences were found in both directions, i.e., for some of these tests, the test–retest reliability was higher for the ISC tests, while for one (i.e., RAVLT Delayed Recall), it was higher for the analog test. For the Clock Test, comparably low test–retest reliability coefficients were found for the analog test versions in previous research, albeit with larger retest intervals which are known to decrease test–retest reliabilities (Hammers et al., 2022). It has been suggested that the subjectivity of the scoring criteria for visual copy and drawing tests as well as the small scoring ranges may contribute to their relatively low test–retest reliability (Kiselica et al., 2020). It is therefore plausible that an automated scoring algorithm based on machine learning may have contributed to an increase in the test–retest reliability of the ISC Clock Test. We also found a higher test–retest reliability for the ISC version of the Star Cancellation Test total score. Another explanation for the difference in test–retest reliability between the digital and analog version of the Clock Test is the difference in scales used for scoring with less of a ceiling effect in the digital Clock Test. It is possible that ceiling effects contributed to a lower test–retest reliability for the Star Cancellation total score and that this effect was somewhat more pronounced in the analog group. The difference in test–retest reliability coefficients for the Category Fluency Test Animals is most likely due to the fact that there was a large number of missing values for the Category Fluency Test for the analog group. This is supported by the finding that for all trials of the Category Fluency Test the test–retest reliability is higher for the ISC than the analog versions. We found a higher test–retest reliability for the analog version of the RAVLT Delayed Recall than the ISC version of the test. A somewhat larger practice effect for the RAVLT Delayed Recall in the ISC group may have contributed to the lower test–retest reliability for the ISC version than the analog version of the test.

Comparing test performance from visit 1 to visit 2 directly via raw scores, we found that differences between visits were mostly comparable between administration modalities. Between group differences were only found for the Clock Test Copy and the Star Cancellation Test duration per correct cancellation with larger practice effects in the ISC group than the analog group. It is important to note that after Bonferroni correction with an alpha of .002, none of the between group differences reached statistical significance. It seems unlikely that differences in administration modality led to a difference in the magnitude of practice effects. It is possible that a larger practice effect was found in the ISC group because of the lower average age in this group. A link between larger practice effects and younger age has been documented in previous research (Calamia, Markon, & Tranel, 2012).

A difference between digital and analog administration mode that is worth noting is the finding that completion times for both the TMT A and the TMT B were higher in the ISC group than the analog group. It seems that this difference is most likely due to an inherent difference in test administration in the digital version of the test. However, despite this difference, the relative difference between visits did not differ between the two groups. In fact, practice effects found for the digital TMT were

comparable to practice effects found in previous research with the paper-pencil version of the TMT in healthy older adults over 1-week retest intervals (Duff, 2014).

Besides establishing the test–retest reliabilities of the different ISC tests, we provide the outcomes of two different methods to calculate reliable change. In clinical practice, it is important for clinicians to be able to determine whether an observed change between assessments is due to an actual change in ability or merely reflects measurement error or a practice effect. For the $R_{CI_{\text{Chelune}}}$ method, 90% and 95% confidence intervals are provided that can inform clinicians on how to interpret observed change. In addition, SRB equations are provided that similarly indicate whether a change found between two assessments is a true change. Consistent with previous research, baseline test performance, retest interval, age, and education predicted test performance at visit 2 with baseline test performance being the strongest predictor for all outcome measures (Duff, 2014; Hammers et al., 2021, 2022). In line with previous research, the variance explained in the different regression models varied per outcome measure (13% to 75%) and was comparable to the results of previous research (Duff, 2014; Hammers et al., 2022; Kiselica et al., 2020). Previous research using a 1-week retest interval in a sample of healthy older adults had very similar regression models as our research for test that overlapped, i.e. the TMT A and B, Letter Fluency Test, and the Category Fluency Test Animals (Duff, 2014). In line with previous research, primarily visuospatial tests such as the Clock Test and the ROCFT had lower predictive capacities (Duff, 2014; Hammers et al., 2022; Kiselica et al., 2020). In contrast to this prior research, test–retest reliabilities found for these tests were higher in the present study which may be partially explained by shorter retest intervals but could also be due to the fact that scoring of the tests was automated.

It is important to keep in mind that mean retest interval in the ISC group was 18 days with a range from 13 to 41 days. In addition, the study included only one repeat visit. In clinical practice, intervals between assessments are often longer and patients may be tested on more than two occasions. Test–retest reliabilities for the ISC tests might be somewhat lower with longer retest intervals, while practice effects might be somewhat smaller over longer retest intervals (Calamia et al., 2012). Although longer retest intervals may only marginally affect RCIs, it would be interesting to explore changes in test–retest reliabilities of ISC tests over longer retest intervals and across several visits. Another limitation of the present study is that it included only healthy adults and exclusion criteria were based solely on self-reported medical history. Test–retest reliabilities may differ for younger age groups and for people with cognitive impairment. In samples with a larger range of baseline and follow-up assessment performances, the variance explained by the regression equations may be higher than in this study which would lead to more accurate predictions of reliable change. Future research should explore to what degree the findings of the present study generalize to more diverse samples with regards to demographic variables, medical comorbidities, and cognitive impairment. In addition, future research should establish the minimal clinically significant differences, i.e., which score changes correspond to real changes in clinical populations that patients and/or relatives consider important.

It is important to mention that the ISC platform has been further developed since the above-described studies have been conducted. The current version of ISC includes additional tests besides the tests included in this paper such as the Symbol Number Matching Test (SNMT). Preliminary analyses show excellent test-retest reliability for the SNMT. Detailed analyses for the SNMT and other tests implemented in future versions of the ISC platform will be published elsewhere.

Conclusion

Test-retest reliabilities for the ISC cognitive tests were found to range between moderate and excellent and were comparable to test-retest reliabilities of the paper-pencil versions of the same tests. The results add to the evidence on the psychometric properties of ISC and support the clinical utility of ISC. Two methods to calculate reliable change are provided, which can be used by clinicians to identify change in cognitive functioning over time. Accurate detection of cognitive change is of paramount importance for different reasons, including aiding in the diagnosis and prognosis of neurological and neurodegenerative diseases over time, and monitoring the effects of treatments on cognition.

Notes

$$1. \frac{((T2 - T1) - (M2 - M1))}{\sqrt{((SD1^2 + SD2^2) * (1 - r))}}$$

with T2 = posttest score, T1 = pretest score, M2 = mean posttest score of group, M1 = mean pretest score of group, SD1 = standard deviation of pretest score, SD2 = standard deviation of posttest score, r = ICC.

- For illustration, also a calculation for one of the regression equations that require back-transformation is provided: the participant had a score of 88.7 on the TMT B on visit 1 and a score of 73.1 on visit 2. The predicted visit 2 score is $10^{(1 + (.004 \times 75) + (-.01 \times 14) + (.001 \times 14) + (.44 \times \log_{10}(88.7))} = 107.42$. The SRB change score is $(73.1 - 107.42)/33.52 = -1.02$. This change score falls within ± 1.645 (90% confidence interval) and therefore indicates that the participant was stable from visit 1 to visit 2. Similarly, when looking at table 4 which provides the RCI for the RCI_{Chelune} method, a difference score of -15.5 on the TMT B falls within the 90% confidence interval ranging from -57.1 to +82.8 indicating that the participant was stable.

Acknowledgements

Laura Klaming, Mandy Spaltman, Stefan Vermeent, Gijs van Elswijk, and Ben Schmand are or have been employed by Philips. Justin B. Miller received consultation fees from Philips. This work was not supported by any grants. The authors would like to thank the Digital Cognitive Diagnostics team at Philips Healthcare for their invaluable work in the development of the ISC tests as well as their practical support during data collection.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

Stefan Vermeent  <http://orcid.org/0000-0002-9595-5373>

Justin B. Miller  <http://orcid.org/0000-0002-4439-6604>

References

- Arrioux, J. P., Cole, W. R., & Ahrens, A. P. (2017). A review of the validity of computerized neurocognitive assessment tools in mild traumatic brain injury assessment. *Concussion (London, England)*, 2(1), CNC31. <https://doi.org/10.2217/cnc-2016-0021>
- Bauer, R. M., Iverson, G. L., Cernich, A. N., Binder, L. M., Ruff, R. M., & Naugle, R. I. (2012). Computerized neuropsychological assessment devices: Joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. *The Clinical Neuropsychologist*, 26(2), 177–196. <https://doi.org/10.1080/13854046.2012.663001>
- Benton, A. I., & Hamsher, K. (1989). *Multilingual Aphasia Examination*. AJA Associates.
- 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. <https://doi.org/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. <https://doi.org/10.1080/13854046.2013.809795>
- Chan, J. Y., Yau, S. T., Kwok, T. C. Y., & Tsoi, K. K. (2021). Diagnostic performance of digital cognitive tests for the identification of MCI and dementia: A systematic review. *Ageing Research Reviews*, 72, 101506. <https://doi.org/10.1016/j.arr.2021.101506>
- Chelune, G., Naugle, R., Lüders, H., Sedlak, J., & Awad, I. (1993). Individual change following epilepsy surgery: Practice effects and base-rate information. *Neuropsychology*, 7(1), 41–52. <https://doi.org/10.1037/0894-4105.7.1.41>
- Cole, W. R., Arrioux, J. P., Schwab, K., Ivins, B. J., Qashu, F. M., & Lewis, S. C. (2013). Test-retest reliability of four computerized neurocognitive assessment tools in an active duty military population. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, 28(7), 732–742. <https://doi.org/10.1093/arclin/act040>
- Duff, K. (2014). One-week practice effects in older adults: Tools for assessing cognitive change. *The Clinical Neuropsychologist*, 28(5), 714–725. <https://doi.org/10.1080/13854046.2014.920923>
- Feenstra, H. E. M., Vermeulen, I. E., Murre, J. M. J., & Schagen, S. B. (2017). Online cognition: Factors facilitating reliable online neuropsychological test results. *The Clinical Neuropsychologist*, 31(1), 59–84. <https://doi.org/10.1080/13854046.2016.1190405>
- Folstein, M. F., Folstein, S. E., White, T., & Messer, M. A. (2010). *Mini-Mental State Examination* (2nd ed). Psychological Assessment Resources.
- Food and Drug Administration. (2020). FDA Medical Devices, 21 C.F.R. SS 882.1470, 2020. Retrieved from <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=882.1470>
- Germine, L., Reinecke, K., & Chaytor, N. S. (2019). Digital neuropsychology: Challenges and opportunities at the intersection of science and software. *The Clinical Neuropsychologist*, 33(2), 271–286. <https://doi.org/10.1080/13854046.2018.1535662>
- Gualtieri, C. T., & Johnson, L. G. (2006). Reliability and validity of a computerized neurocognitive test battery, CNS vital signs. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, 21(7), 623–643. <https://doi.org/10.1016/j.acn.2006.05.007>

- Hammers, D. B., Kostadinova, R., Unverzagt, F. W., & Apostolova, L. G. for the Alzheimer's Disease Neuroimaging Initiative. (2022). Assessing and validating reliable change across ADNI protocols. *Journal of Clinical and Experimental Neuropsychology*, 44(2), 85–102. <https://doi.org/10.1080/13803395.2022.2082386>
- Hammers, D. B., Suhrie, K. R., Dixon, A., Porter, S., & Duff, K. (2021). Reliable change in cognition over 1 week in community-dwelling older adults: A validation and extension study. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, 36(3), 347–358. <https://doi.org/10.1093/arclin/acz076>
- Hinton-Bayre, A. (2010). Deriving reliable change statistics from test-retest normative data: Comparison of models and mathematical expressions. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, 25(3), 244–256. <https://doi.org/10.1093/arclin/acq008>
- 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. <https://doi.org/10.1037//0022-006x.59.1.12>
- Kessels, R. P. C. (2019). Improving precision in neuropsychological assessment: Bridging the gap between classic analog tests and paradigms from cognitive neuroscience. *The Clinical Neuropsychologist*, 33(2), 357–368. <https://doi.org/10.1080/13854046.2018.1518489>
- Kiselica, A. M., Kaser, A. N., Webber, T. A., Small, B. J., & Bengel, J. F. (2020). Development and preliminary validation of standardized regression-based change scores as measures of transitional cognitive decline. *Archives of Clinical Neuropsychology*, 35(7), 1168–1181. <https://doi.org/10.1093/arclin/aca042>
- Koo, T. K., & Li, M. Y. (2017). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2004). *Neuropsychological assessment* (4th ed.). Oxford University Press.
- Littleton, A. C., Register-Mihalik, J. K., & Guskiewicz, K. M. (2015). Test-retest reliability of a computerized concussion test: CNS vital signs. *Sports Health*, 7(5), 443–447. <https://doi.org/10.1177/1941738115586997>
- Manos, P. J., & Wu, R. (1994). The ten point clock test: A quick screen and grading method for cognitive impairment in medical and surgical patients. *International Journal of Psychiatry in Medicine*, 24(3), 229–244. <https://doi.org/10.2190/5A0F-936P-VG8N-0F5R>
- McSweeney, A. J., Naugle, R. I., Chelune, G. J., & Lüders, H. (1993). T-scores for change: An illustration of a regression approach to depicting change in clinical neuropsychology. *Clinical Neuropsychologist*, 7(3), 300–312. <https://doi.org/10.1080/13854049308401901>
- Meyers, J. E., & Meyers, K. R. (1995). *Rey complex figure test and recognition trial*. Psychological Assessment Resources.
- Miller, J. B., & Barr, W. B. (2017). The technology crisis in neuropsychology. *Archives of Clinical Neuropsychology*, 32(5), 541–554. <https://doi.org/10.1093/arclin/acx050>
- Nakayama, Y., Covassin, T., Schatz, P., Nogle, S., & Kovan, J. (2014). Examination of the test-retest reliability of a computerized neurocognitive test battery. *The American Journal of Sports Medicine*, 42(8), 2000–2005. <https://doi.org/10.1177/0363546514535901>
- O'Brien, A. M., Bartlett, A. N., Frost, N., & Casey, J. E. (2022). Inflated scaled scores on the digital WISC-V coding subtest in a Canadian sample. *Applied Neuropsychology: Child*, 11(2), 150–157. <https://doi.org/10.1080/21622965.2020.1773270>
- Rabin, L. A., Paolillo, E., & Barr, W. B. (2016). Stability in test-usage practices of clinical neuropsychologists in the United States and Canada over a 10-year period: A follow-up survey of INS and NAN members. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, 31(3), 206–230. <https://doi.org/10.1093/arclin/acw007>
- Rabin, L. A., Spadaccini, A. T., Brodale, D. L., Grant, K. S., Elbulok-Charcape, M. M., & Barr, W. B. (2014). Utilization rates of computerized tests and test batteries among clinical neuropsychologists in the United States and Canada. *Professional Psychology: Research and Practice*, 45(5), 368–377. <https://doi.org/10.1037/a0037987>

- Riordan, P., Lombardo, T., & Schulenberg, S. E. (2013). Evaluation of a computer-based administration of the Rey Complex Figure Test. *Applied Neuropsychology Adult*, 20(3), 169–178. <https://doi.org/10.1080/09084282.2012.670171>
- Schmand, B. (2019). Why are neuropsychologists so reluctant to embrace modern assessment techniques? *The Clinical Neuropsychologist*, 33(2), 209–219. <https://doi.org/10.1080/13854046.2018.1523468>
- Schmand, B., Groenink, S. C., & van den Dungen, M. (2008). Letter fluency: Psychometrische eigenschappen en Nederlandse normen [Letter Fluency: Psychometric characteristics and Dutch norms]. *Tijdschrift Voor Gerontologie en Geriatrie*, 39(2), 64–76. <https://doi.org/10.1007/BF03078128>
- Schmidt, M. (1996). *Rey auditory verbal learning test: A handbook*. Western Psychological Services.
- Staffaroni, A. M., Tsoy, E., Taylor, J., Boxer, A. L., & Possin, K. L. (2020). Digital cognitive assessments for dementia. *Practical Neurology (Fort Washington, Pa.)*, 20(20), 24–45.
- Strauss, E., Sherman, E. M., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary*. American Chemical Society. U.S. Census Bureau (2017). Current Population Survey, Annual Social and Economic Supplement.
- Vermeent, S., Spaltman, M., Van Elswijk, G., Miller, J. B., & Schmand, B. (2022). Philips IntelliSpace Cognition digital test battery: Equivalence and measurement invariance compared to traditional analog test versions. *The Clinical Neuropsychologist*, 36(8), 2278–2299. <https://doi.org/10.1080/13854046.2021.1974565>
- US Census Bureau. (2017). <https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/2017/>
- Weintraub, S., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Zelazo, P. D., Bauer, P. J., Carlozzi, N. E., Slotkin, J., Blitz, D., Wallner-Allen, K., Fox, N. A., Beaumont, J. L., Mungas, D., Nowinski, C. J., Richler, J., Deocampo, J. A., Anderson, J. E., Manly, J. J., Borosh, B., ... Gershon, R. C. (2013). Cognition assessment using the NIH Toolbox. *Neurology*, 80(11 Suppl 3), S54–S64. <https://doi.org/10.1212/WNL.0b013e3182872ded>
- Wilson, B., Cockburn, J., & Halligan, P. (1987). *Behavioral inattention test manual*. Thames Valley Test Company.
- Zygouris, S., & Tsolaki, M. (2015). Computerized cognitive testing for older adults: A review. *American Journal of Alzheimer's Disease and Other Dementias*, 30(1), 13–28. <https://doi.org/10.1177/1533317514522852>