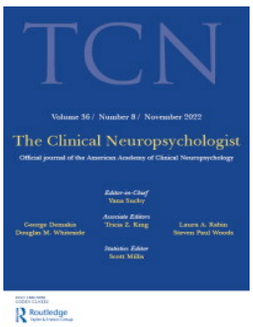

The Clinical Neuropsychologist

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ntcn20

Philips IntelliSpace Cognition digital test battery: Equivalence and measurement invariance compared to traditional analog test versions

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

To cite this article: Stefan Vermeent, Mandy Spaltman, Gijs van Elswijk, Justin B. Miller & Ben Schmand (2022) Philips IntelliSpace Cognition digital test battery: Equivalence and measurement invariance compared to traditional analog test versions, The Clinical Neuropsychologist, 36:8, 2278-2299, DOI: 10.1080/13854046.2021.1974565

To link to this article: https://doi.org/10.1080/13854046.2021.1974565

THE CLINICAL NEUROPSYCHOLOGIST
2022, VOL. 36, NO. 8, 2278–2299
https://doi.org/10.1080/13854046.2021.1974565

Routledge
Taylor & Francis Group

Philips IntelliSpace Cognition digital test battery: Equivalence and measurement invariance compared to traditional analog test versions

Stefan Vermeent[a], Mandy Spaltman[a], Gijs van Elswijk[a], Justin B. Miller[b] and Ben Schmand[a]

[a]Digital Cognitive Diagnostics, Philips Healthcare, Eindhoven, The Netherlands; [b]Cleveland Clinic Lou Ruvo Center for Brain Health, Las Vegas, NV

ABSTRACT

Objective: To collect evidence of validity for a selection of digital tests on the Philips IntelliSpace Cognition (ISC) platform.

Method: A total of 200 healthy participants (age 50-80) completed both the ISC battery and an analog version of the battery during separate visits. The battery included the following screeners and cognitive tests: Mini-Mental State Examination (2nd edition), Clock Drawing Test, Trail-Making Test (TMT), Rey Auditory Verbal Learning Test (RAVLT), Rey-Osterrieth Complex Figure Test (ROCFT), Letter Fluency, Star Cancellation Test, and Digit Span Test. The ISC tests were administered on an iPad Pro and were automatically scored using designated algorithms. The analog tests were administered in line with existing guidelines and scored by trained neuropsychologists. Criterion validity was established through relative agreement coefficients and raw score equivalence tests. In addition,measurement invariance analysis was used to compare the factor structures of both versions. Finally, we explored effects of demographics and experience with digital devices on performance.

Results: We found fair to excellent relative agreement between test versions. Absolute equivalence was found for RAVLT, Letter Fluency, Star Cancellation Test, and Digit Span Test. Importantly, we demonstrated equal loadings of the digital and analog test versions on the same set of underlying cognitive domains. Demographic effects were mostly comparable between modalities, and people's experience with digital devices was found to only influence performance on TMT B.

Conclusions: This study provides several sources of evidence for the validity of the ISC test battery, offering an important step in validating ISC for clinical use.

ARTICLE HISTORY
Received 25 February 2021
Accepted 26 August 2021

KEYWORDS
neuropsychology; psychological tests; cognition; digital technology; automation; measurement invariance

Digital cognitive testing has been gaining traction in recent years as a powerful and flexible alternative to traditional analog testing. Its potential benefits are well-documented, and include a high level of standardization, a significant increase

CONTACT Gijs van Elswijk gijs.van.elswijk@philips.com Digital Cognitive Diagnostics, Philips Healthcare, High Tech Campus 34, AE Eindhoven, 5656, The Netherlands
*Stefan Vermeent is now at Department of Psychology, Utrecht University, The Netherlands.
Supplemental data for this article is available online at https://doi.org/10.1080/13854046.2021.1974565.

in the amount of performance information, and a more efficient workflow both in terms of administration and scoring (Kessels, 2019; Riordan et al., 2013; Zygouris & Tsolaki, 2015). IntelliSpace Cognition (ISC) developed by Philips is a digital adaptation of standardized cognitive tests, which is intended for use in clinical settings. ISC contains digital versions of existing analog tests (as opposed to tests specifically created for the digital domain; Koo & Vizer, 2019). In a previous article, we provided initial validation data using a prototype version of ISC in a Dutch sample and concluded that the digital tests measured the same underlying constructs as commonly found in the (analog-based) literature (Vermeent et al., 2020).

In the present article, we sought to extend these findings by providing additional validation data directly comparing the clinically available version of the ISC battery (second version) against the analog versions of the tests in a representative US sample. Such direct comparisons between digital and analog test versions are crucial because despite their relative equivalence on the surface, digital tests might differ in various ways from their analog counterparts (Bauer et al., 2012; American Educational Research Association et al., 2014), and thus require dedicated psychometric validation. While the prototype version of ISC presented in Vermeent and colleagues (2020) still required substantial involvement of a clinician (i.e., providing instructions, correcting errors, rating performance), the clinically available version of the ISC battery presented here contains tests that only required minimal supervision.

Converting analog tests to a digital format can have subtle but important effects on performance. Relevant factors that might lead to differences between modalities are screen size (Grinspoon, Passell, Scheuer, Ressler, & Germine et al., 2019), platform-specific hardware or software aspects (Dadey et al., 2018; Passell et al., 2020), and input type (e.g., computer mouse vs. finger vs. digital pen; Germine et al., 2019). Such differences are inevitable to a degree. For example, performing a drawing test on a tablet using a digital pen provides a different user experience from performing that same test on paper (Spreij et al., 2020). When differences between modalities are systematic and relatively consistent between patients, they are not necessarily problematic, given that separate norms are provided (Bauer et al., 2012). Problems arise, however, when changes introduced in a digital assessment alter the cognitive construct that is being measured. Strong arguments have been made that digital adaptations of cognitive tests, just as any other new cognitive test, should be submitted to scrutinous validation and reliability studies (Bauer et al., 2012; American Educational Research Association et al., 2014). However, evidence for validity and reliability is still lacking for many digital tests (Schlegel & Gilliland, 2007; Schmand, 2019; Wild et al., 2008), and existing evidence is often mixed (e.g., Arrieux et al., 2017; Bailey et al., 2018; Björngrim et al., 2019; Cole et al., 2018; Daniel, Wahlstrom, & Zhang, 2014; Gualtieri & Johnson, 2006; Morrison et al., 2018).

Most cognitive tests are to some extent influenced by demographic characteristics of the test-taker such as age, education level, and sex (Strauss et al., 2006). Digital test versions are no different, though they can also potentially introduce unique demographic effects that influence performance (Germine et al., 2019). For example, older adults tend to have less computer experience and generally perceive digital technology as more difficult to use than younger adults (Hauk et al., 2018; Lee Meeuw Kjoe et al., 2021; Rabipour & Davidson, 2020; Wild et al., 2012), which may make a digital test

battery more challenging or intimidating for older adults (Rabipour & Davidson, 2020; Zygouris & Tsolaki, 2015). Any such differences caused by level of familiarity might very well soon or already be outdated given the speed with which digital technologies continue to pervade most aspects of society for all age groups (Hülür & MacDonald, 2020; Miller & Barr, 2017). However, they must nevertheless be taken into consideration, seeing as how recent estimates of ownership of handheld devices is only around 40% for tablets and nearly 50% for smartphones in households over 65 years of age (Ryan, 2018). These numbers remain low relative to younger age groups despite constituting roughly a 10% increase in three years' time (File & Ryan, 2014).

The current article investigates several aspects of validity of the ISC digital cognitive test battery using a large, representative group of healthy older adults who completed both the clinical version of the ISC battery (second edition) and an analog battery containing the same tests across two visits. First, we provide evidence of criterion validity by examining the agreement between scores on the ISC test versions and scores on a comparative test battery consisting of analog versions of the same tests. Given our previous work with a prototype version of ISC, we hypothesized that performance data generated by ISC would show an acceptable level of relative agreement with analog counterparts (as assessed through intraclass correlations) as well as absolute equivalence on a raw scale. Second, we provide additional evidence of construct validity via measurement invariance analyses to test whether performance on both versions is related to the same underlying cognitive domains. We hypothesized that the same factor model provides good fit to the data generated by both the ISC and the analog battery and that tests of both versions measure the same latent constructs. Third, we explored the effect of demographic characteristics and experience with digital devices on performance for both modalities.

Materials and methods

Participants

A total of 200 participants were recruited through a research recruitment agency, following a sampling plan that stratified for sex, education level and racial/ethnic background according to the U.S. demographic (U.S. Census Bureau, 2017) within predefined age groups spanning 5 years each (see Table 1) . Participants were healthy volunteers (Mean MMSE score = 27.7 ($SD = 1.8$), range = 21 − 30) between the ages of 50 and 80 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 day of testing, as well as normal fine and gross motor ability. The exclusion criteria were self-reported presence of neurological, autoimmune, major mental illness, mood disorder, anxiety and/or autism spectrum disorders; intellectual disability; recent functional changes in reference to activities of daily living (ADLs); substance abuse or dependence; aphasia or any cognitive difficulties for which participants were seeking medical help at that time; admission to a hospital, or residence in an assisted living facility, nursing home or psychiatric facility at the time of the visit; history of head trauma with loss of consciousness greater than 20 minutes; history of resuscitation, exposure to ECT, radiation to the central nervous system;

Table 1. Participant Characteristics.

	ISC, then Analog ($n = 99$)	Analog, then ISC ($n = 101$)	Test statistic	p
Age, mean (*SD*)	67.2 (7.0)	66.3 (7.5)	$t(197.38) = 0.88$	.383
Years of Education, mean (*SD*)	13.7 (2.2)	13.6 (2.3)	$t(197.49) = 0.14$	.893
Sex, % female	50.1	52.5	$\chi^2(1) = 0.02$	.891
Race			$\chi^2(3) = 3.95$	.267
White, %	80.8	71.3		
Black, %	11.1	12.9		
Hispanic, %	6.1	8.9		
Other, %	2.0	6.9		
MMSE-2 score, median[a, b]	28	28	U = 4909.5	.824
Clock Drawing Test, median[a, b]	8	9	U = 4928.5	.960
Study location			$\chi^2(3) = 4.24$	.236
Totowa (NJ), %	29.3	34.7		
Orlando (FL),%	18.2	25.7		
Philadelphia (PA), %	28.3	24.8		
Sacramento (CA), %	24.2	14.9		
Test-retest interval in days, mean (*SD*)	19.9 (10.1)	18.1 (6.4)	$t(165.52) = 1.53$	.129

Note. ISC = IntelliSpace Cognition; MMSE-2 = Mini-Mental State Examination 2nd Edition.
[a]Scoring according to Manos and Wu (1994; range 0-10).
[b]Non-parametric Mann–Whitney test.

previous experience with undergoing neuropsychological testing; use of medication that might impact test performance (e.g., anticonvulsants, benzodiazepines); received chemotherapy treatment in the last 2 months. Recruitment and testing took place at four locations in the United States of America: Totowa (NJ), Orlando (FL), Philadelphia (PA), and Sacramento (CA).

After data collection was completed, a total of 11 ISC test scores were excluded casewise due to technical difficulties, retaining the other test scores for analyses: Three Rey Auditory Verbal Learning (RAVLT) learning trial scores were missing due to missing audio recordings; one Letter Fluency Test score was missing due to a missing audio recording; one Rey-Osterrieth Complex Figure Test (ROCFT) Immediate Recall score was missing because the drawing was not recorded due to a human error; and one Trail Making Test (TMT) A score was missing due to a technical error. Four participants did not reach the final target of the TMT B (i.e., the participant stopped the test before reaching the area of the final target). In addition, one participant only partially completed the Star Cancellation Test, crossing out only 30 of the 54 targets. For the latter we were unable to determine whether or not this issue was caused by a technical failure (i.e., digital pen detection failure by the iPad). While the product version of ISC handles non-completion more leniently, we decided to exclude these participants here. Whereas non-completion is likely related to cognitive functioning in a clinical sample, all participants in our sample were healthy and should therefore be able to complete all tests. Finally, five participants had missing data on the familiarity questionnaire and where therefore excluded from the regression analyses involving this measure (see below).

Measures

An overview of the included outcome measures per test, and a description of the administration and scoring differences between the two modalities is presented in Table 2. The tests were administered in a fixed order for every participant and at

Table 2. Overview of tests.

Test	Outcome measures	Administration	Duration[a]	Scoring
MMSE-2 *Screener*	Standard score (0–30)	In accordance with Folstein et al. (2010). Instructions adapted to reflect digital format for ISC test and provided orally by test administrator.	4 minutes	*ISC:* Immediate scoring in-tool by test administrator. Analog: Transcription by test administrator; retrospective scoring by neuropsychologists.
Clock Drawing Test *Screener*	Total score Manos and Wu (1994; 0–10)	In accordance with Strauss et al. (2006). Instructions adapted to reflect digital format for ISC test and provided by standardized built-in audio.	1 minute	*ISC:* Retrospective scoring by neuropsychologists. *Analog:* Retrospective scoring by neuropsychologists.
Clock Drawing Test Copy *Screener*			1 minute	
RAVLT Learning Trials	Total score (0–75)	In accordance with Schmidt (1996). Instructions adapted to reflect digital format for ISC test and provided orally by test administrator. *ISC:* Response captured by audio recording. *Analog:* Response captured by test administrator transcription.	6 minutes	*ISC:* Retrospective scoring by speech-recognition algorithm based on raw audio recordings. *Analog:* Transcription by test administrator; retrospective scoring by neuropsychologists.
RAVLT Interference	Total score (0–15)		1 minute	
RAVLT Immediate Recall			1 minute	
RAVLT Delayed Recall			1 minute	
TMT A	Duration (in seconds)	In accordance with Strauss et al. (2006). Instructions adapted to reflect digital format for ISC test and provided by standardized built-in audio.	2 minutes	*ISC:* Immediate scoring by algorithm. *Analog:* Immediate timing by test administrator (using stopwatch).
TMT B			3 minutes	
Star Cancellation Test	Total score (0–54), Duration per Correct Cancellation (in seconds)	In accordance with Wilson et al. (1987). Instructions adapted to reflect digital format for ISC test and provided orally by test administrator.	1 minute	*ISC:* Immediate scoring by algorithm. *Analog:* Immediate timing by test administrator (using stopwatch); retrospective scoring by neuropsychologists.
ROCFT Copy	Total score (0–36)	Instructions adapted to reflect digital format for ISC test and provided by standardized built-in audio.	3 minutes	*ISC:* Retrospective scoring by algorithm (without explicit scoring rules). *Analog:* Retrospective scoring by neuropsychologists according to Meyers and Meyers (1995).
ROCFT Immediate Recall			2 minutes	
Letter Fluency Test	Total score	In accordance with Benton and Hamsher (1989) and Schmand, Groenink, and van den Dungen (2008). Instructions adapted to reflect digital format for ISC test and provided orally by test administrator. *ISC:* Response captured by audio recording. *Analog:* Response captured by test administrator transcription.	5 minutes	*ISC:* Retrospective scoring by speech-recognition algorithm based on audio files. *Analog:* Transcription by test administrator; retrospective scoring by neuropsychologists.
Digit Span Forward	Total score (0–12)	In accordance with Lezak, Howieson, and Bigler (2012). Instructions adapted to reflect digital format for ISC test 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.	9 minutes	*ISC:* Immediate scoring by algorithm. *Analog:*Transcription by test administrator; retrospective scoring by neuropsychologists.
Digit Span Backward				

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

[a]Average duration of all ISC assessments rounded to the closest whole minute.

every visit: Mini Mental State Examination (2nd edition), RAVLT Learning Trials, RAVLT Interference, RAVLT Immediate Recall, TMT A, TMT B, O-Cancellation test, Clock Drawing Test, Clock Drawing Test Copy Trial, Star Cancellation Test, RAVLT Delayed Recall, RAVLT Recognition, ROCFT Copy Trial, Letter Fluency Test, ROCFT Immediate Recall, Digit Span Forward, Digit Span Backward, and Category Fluency Test. The decision to fix the test order ensured that several 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; no verbal tests between the different RAVLT trials and no drawing tests between the ROCFT trials). The analog versions of the tests were administered in the same order, according to standardized administration rules as used in clinical practice. We did not use parallel versions of the tests across assessments for two reasons: First, using parallel versions would make it more challenging to draw conclusions about equivalence, especially in the case of deviations between modalities. Second, for some of the tests included in the ISC battery no parallel versions exist in the literature, and not all parallel versions are as well-validated as the main versions.

Several designated algorithms were developed for ISC tests that required scoring following fixed scoring rules. The ROCFT Copy and Immediate Recall drawings were scored by a deep learning algorithm and the audio recordings of the RAVLT and Letter Fluency trials were scored by speech recognition algorithms. The technical details about the algorithms and the reliability analysis are beyond the scope of this paper. In short, in order to assess their reliability relative to human raters, we randomly sampled subsets of ROCFT drawings and audio recordings of the RAVLT and Letter Fluency to be scored by human raters. These subsets were sampled from a larger pool of participants than discussed in this paper, some of whom only completed the digital battery and are therefore not included for the main analyses described in this article. Within these subgroups, we calculated inter-rater reliabilities among the human raters and compared them to the human-algorithm inter-rater reliabilities (see Table 3). Although the human-algorithm reliabilities tended to be slightly lower than the reliabilities among human raters, they were all reliably above .90, indicating strong agreement between the human raters and the automated algorithms.

The ISC tests were implemented on the Philips ISC platform to be similar to these analog versions and were performed using an iPad Pro tablet (3rd generation) running on iOS 12 with a screen size of 12.9 inches and resolution of 2732×2048 pixels. Drawings were made using an Apple Pencil for the iPad Pro. For tests that required a verbal response, audio recordings were made using the iPad's internal microphone. Two tests, namely O-Cancellation Test and Category Fluency Test, as well as the Recognition trial of the RAVLT have not been implemented in the second clinical release of ISC, which means that the automated algorithms necessary to score these tests were not yet developed at the time of writing. Because of this, these tests are not included in this article.

Prior to cognitive testing, participants completed a digital familiarity questionnaire specifically developed for the purpose of this study about their everyday experience with various digital devices. The questionnaire consisted of six questions, asking about frequency of use of digital devices ("Please indicate how often you use the following devices: Desktop PC/laptop PC/smartphone/tablet") and self-ratings of use (e.g., "how

Table 3. Inter-Rater Reliabilities between Human Raters and Algorithm Scores.

				Agreement (Intraclass correlation) [95% confidence interval]	
Test	Measure	*N* trials	*N* raters	Human raters	Human raters-Algorithms
ROCFT	Drawing score	200	6	.97 [.97, .98]	.93 [.92, .95]
Letter Fluency	F + A + S total score	3 * 100	3	.99 [.98, .99]	.96 [.94, .97]
RAVLT	Word list A trials total score	7 * 100	4	.99 [.99, .99]	.97 [.95, .98]

Note. RAVLT = Rey Auditory Verbal Learning Test; ROCFT = Rey-Osterrieth Complex Figure Test.

would you rate yourself on the use of a smartphone/tablet?"). Here we focus on the subset of four items probing the frequency of use. Participants answered using a four-point scale ranging from "Not" to "(almost) daily". An overall measure of experience with digital devices was computed by calculating the sum of these four measures, with possible scores ranging from 0 to 12.

Procedure

Ethics approval for the study was obtained from the Internal Committee of Biomedical Experiments within Philips as well as from an external institutional review board (Western IRB). The study was conducted according to the principles of the Declaration of Helsinki and per the requirements of the ISO 14155 standard (Good Clinical Practice). The study was registered on ClinicalTrials.gov (NCT03801382). Continuous quality assurance took place during data collection by inspecting the ISC data for completeness and potential test order deviations, through regular team meetings to evaluate and reflect on the process, and through monitoring activities of the study by an independent clinical research organization.

See Figure 1 for a flowchart of each visit. Upon inclusion, participants were assigned randomly to one of two groups with minimization for demographic characteristics. Half of the participants ($n = 99$) first received the ISC tests, and the other half ($n = 101$) received the analog measures first. After an interval of at least 14 days, participants returned to the clinic to complete the other test modality. The total duration per visit

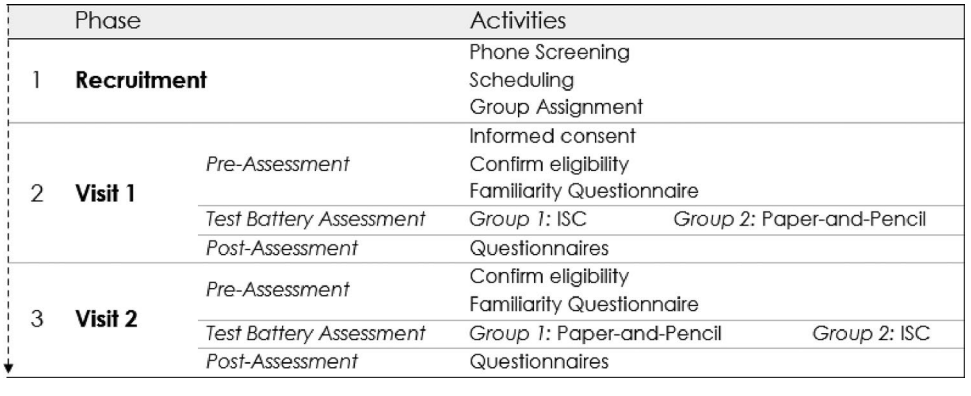

Figure 1 . Study visit flow chart.

was approximately 1.5 hours. At the beginning of the visit, the participants signed an informed consent form (only at visit 1), the test administrator confirmed eligibility of the participant by verifying all inclusion- and exclusion criteria, and participants completed a digital familiarity questionnaire. Next, participants completed either the ISC or analog version of the test battery. Finally, participants filled out usability questionnaires. The visits took place in a dedicated examination room at one of the four study locations. Sources of lighting reflections on the iPad (from a window or a lamp place right above) were eliminated. Four test administrators (one per study location) were hired and trained on the use of ISC and the study protocol. The test administrators were licensed psychologists with clinical experience in neuropsychological testing. For each individual participant all tests were administered by the same test administrator. Participants entered the study voluntarily and received a 75 US dollar compensation for each visit. Participants did not receive feedback about their performance on the tests.

Analyses

All analyses were done using R 4.0 (R Core Team, 2020).

Cross-modal agreement and equivalence. Each ISC test was compared with its analog counterpart through intraclass correlations (ICC) for consistency using two-way random effects models. The ICC values were interpreted using the guidelines described by Cicchetti (1994): <0.40 indicating poor agreement; 0.40 – 0.59 indicating fair agreement; 0.60 – 0.74 indicating good agreement, and >.75 indicating excellent agreement.

In addition, we directly tested for statistical equivalence between analog and ISC test scores using the two one-sided *t*-tests procedure (TOST; Schuirmann, 1987; Lakens et al., 2018) using the TOSTER package (Lakens, 2017). In TOST, one specifies a range within which the difference between two values will be regarded as equivalent. Through two one-tailed t-tests, the TOST addresses the composite null-hypothesis that the observed difference either exceeds the lower or exceeds the upper boundary value of the equivalence range. Equivalence is established when both tests are statistically significant (i.e., both null hypotheses can be rejected). Since equivalence requires both tests to be significant, it is not necessary to control for multiple comparisons (Lakens et al., 2018). Thus, equivalence can be inferred through a statistically significant effect rather than by failure to detect a significant difference. Note that a significant TOST result does not necessarily mean that the two scores are perfectly identical. Rather, it means that any difference between the scores is too small to constitute a relevant effect (Lakens et al., 2018).

To test for equivalence between the ISC and analog test scores, we defined practical equivalence by means of lower and upper equivalence bounds of −0.3 *SD* and +0.3 *SD* of the scores of the analog test versions. These bounds corresponded to the standard error of measurement of a reliable test (.91). The TOST was supplemented by paired samples t-tests comparing the raw means of the ISC and analog test scores. The effect sizes of these t-tests were assessed through Cohen's *d*, using 0.2, 0.5 and

0.8 as cut-offs for small, medium and large effects, respectively. The combination of *t*-test and TOST analyses informed us whether 1) there were mean differences between test scores and if so, 2) whether these fell within or outside a predefined practical equivalence range. As the TOST test provides two *p*-values (one for each equivalence bound), we only report the least extreme *p*-value (Lakens et al., 2018).

Cross-modal measurement invariance. We used longitudinal measurement invariance based on CFA to assess the construct validity of the ISC test battery, that is, whether the same cognitive constructs underlay performance on the ISC test battery compared to the analog test battery (see below for more information on the exact statistical procedure). Employing a longitudinal extension of the standard measurement invariance procedure allowed us to take into account the within-subject design of our study. Typically, longitudinal invariance is used to assess measurement changes in the same tests over time (Mackinnon et al., 2021). Here, however, the interest was not so much in change over time per se but in potential change across modalities. As the order in which participants completed the test versions across sessions was counterbalanced, this meant collapsing across sessions so that all ISC versions and all analog versions were grouped together. Although the principle is identical, this makes "longitudinal" a slight misnomer.

We used CFA as opposed to more exploratory approaches because the factor structure of the analog test versions has been well-established in the literature (e.g., Larrabee, 2015; Lezak et al., 2012; Strauss et al., 2006). This information can be leveraged in a CFA context to apply a-priori constraints to the model, as opposed to exploratory factor analysis (EFA) which assumes no prior knowledge about the data and is more sensitive to patterns that are unique to the sample under investigation. The measurement invariance approach has advantages for establishing construct validity in two ways: First, we started by fitting the model to the analog data, thereby confirming whether its specifications were sensible in the context of the traditional measures. Second, it allows an incremental assessment of the extent to which the ISC tests adhered to the same factor structure, and, in the case of measurement invariance, which statistical dimension would be the cause of this invariance.

The CFA model consisted of four latent cognitive domains (see Figure 2) and was fitted using the *lavaan* package (Rosseel, 2012). As is standard in longitudinal measurement invariance, we specified one model containing both the ISC and analog measures, each with mappings to their own latent factor. Thus, the specification contained two versions of each latent factor (e.g., one "ISC version" and one "analog version" of working memory). The cognitive domains were scaled by fixing their error variances to 1. In order to reduce the number of parameters in the configural model, we fixed residual covariances of the manifest variables to be equal across modalities (Cole & Maxwell, 2003). The model specification is a slight adaptation of a theory-driven model that was previously used to establish initial evidence of validity of a Dutch prototype version of the ISC test battery (Vermeent et al., 2020). The model differed from the one presented in Vermeent et al. because some tests included in the prototype version were not included in the clinically available second version of ISC studied here.

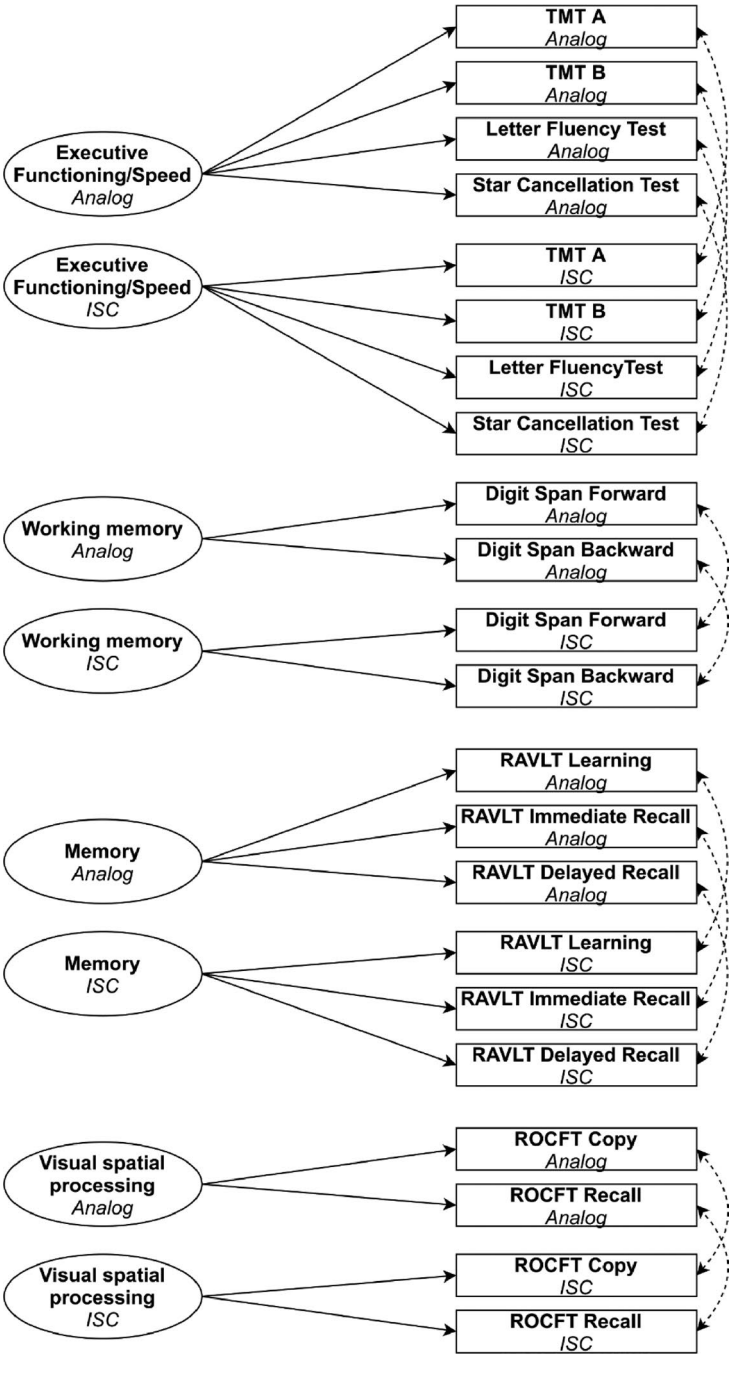

Figure 2 .*Configural* model used for the measurement invariance analysis. Unidirectional arrows represent factor loadings; Bidirectional arrows represent covariances; dashed lines represent constrained parameters (see main text); Ellipses represent latent variables; Rectangles represent manifest variables. Note that in the fitted model, the latent factors are allowed to freely covary with each other. For the sake of clarity, these covariance parameters are omitted from the Figure. TMT = Trail Making Test; RAVLT = Rey Auditory Verbal Learning Test; ROCFT = Rey-Osterrieth Complex Figure Test.

Measurement invariance was assessed in several steps (Vandenberg & Lance, 2000). First, we assessed the equivalence of the underlying covariance matrices between test modalities (i.e., configural equivalence) by fitting the model to the ISC and analog tests without any equivalence constraints (except for the residual covariances between manifest variables as described above). In other words, this step tested whether the specification of the model structure resulted in good fit to the observed data of both test modalities. Model fit was assessed using several quantitative indices, following recommendations by Schermelleh-Engel et al. (2003): a good fit to the data was qualified by a non-significant chi-squared test ($p > .05$), Standardized Root Mean Square Residual (SRMR) ≤ 0.05 (acceptable fit ≤ 0.10); Root Mean Square Error of Approximation (RMSEA) ≤ 0.05 (acceptable fit ≤ 0.08); Comparative Fit Index (CFI) ≥ 0.97 (acceptable fit ≥ 0.95). As the chi-squared test can be sensitive to sample size, we assign relatively more weight to the SRMR, RMSEA and CFI (Schermelleh-Engel et al., 2003).

Next, we fitted a more stringent model version constraining the factor loadings to be equal across modalities (i.e., metric invariance). This model tested whether the tests contributed equally to the latent factors across modalities. Finally, an additional model placed further equality constraints on the test intercepts (i.e., scalar invariance). Equality in test intercepts indicates that the same score on the latent variable results from similar intercepts for the tests or whether there are systematic mean differences (e.g., when an average domain score is linked to faster completion times on one modality compared to the other). Thus, the metric invariance model tested whether the latent variables represented the same underlying construct across domains, while the scalar invariance tested for differences in offset in the relationship between cognitive tests and latent variables. Model fit of consecutive models was compared through chi-squared difference tests. If measurement invariance was not supported for a given model, more constrained models were not considered. In case of non-invariance, we attempted to localize the parameters driving the non-invariance through modification indices.

Effects of demographics and experience with digital devices. We constructed linear regression models per modality with test scores as outcome variables to assess the effects of demographic characteristics, experience with digital devices, and modality order (i.e., analog first or ISC first). Included as predictors were age (in years), education (in years), sex (binary coding, using female as the reference category), experience with digital devices (see *Measures* section) and modality order (binary coding, using analog first as the reference category). We applied Benjamini-Hochberg corrections to *p*-values per test across modality models.

Results

Cross-modal agreement and equivalence

See Table 4 for an overview of agreement and equivalence statistics of all tests. The ICC values between the ISC and analog test versions ranged from .45 for ROCFT Copy total score to .76 for Letter Fluency total score, predominantly showing fair to good agreement (with the exception of Letter Fluency, which showed excellent agreement). Practical equivalence as assessed through the TOST procedure was established for all

Table 4. Comparisons between test scores of intelliSpace Cognition and analog test versions.

	Analog	ISC	Difference Test				TOST		
Test	*M* (*SD*)	*M* (*SD*)	ICC	*t* (df)	*p*	Cohen's *d*	*t* (df)	*p*	90% CI (*SD*s)
TMT A duration (*n* = 199)	34.0 (9.9)	41.5 (13.6)	.52[b]	9.05 (198)	<.001	0.64	−5.44 (198)	>.999	−0.7, −0.5
TMT B duration (*n* = 194)	81.7 (35.7)	105.8 (44.9)	.55[b]	8.70 (193)	<.001	0.62	−4.84 (193)	>.999	−0.7, −0.5
Digit Span Forward total score (*n* = 200)	8.0 (2.0)	7.8 (2.1)	.60[c]	1.81 (199)	.072	0.13	2.81 (199)	.003	0. 0.2
Digit Span Backward total score (*n* = 200)	6.1 (1.9)	6.4 (2.5)	.47[b]	1.73 (199)	.086	0.12	−1.80 (199)	.037	−0.2, 0
Letter Fluency total score trial 1-3 (*n* = 199)	40.9 (11.7)	39.7 (11.3)	.76[d]	2.11 (198)	.036	0.15	4.13 (198)	<.001	0.0, 0.2
ROCFT Copy total score (*n* = 200)	27.0 (6.7)	28.8 (6.7)	.45[b]	3.64 (199)	<.001	0.26	−0.38 (199)	.352	−0.4, −0.1
ROCFT Immediate Recall total score (*n* = 199)	11.4 (5.9)	9.9 (6.6)	.71[c]	4.35 (198)	<.001	0.31	0.91 (198)	.182	0.1, 0.3
RAVLT Learning total score trial 1-5 (*n* = 197)	43.5 (9.6)	44.6 (10.6)	.59[b]	1.65 (196)	.186	0.12	−2.76 (196)	.003	−0.2, 0
RAVLT Immediate Recall total score (*n* = 200)	8.0 (3.2)	8.3 (3.2)	.60[c]	1.45 (199)	.148	0.10	−3.29 (199)	<.001	−0.2, 0
RAVLT Delayed Recall total score (*n* = 200)	7.9 (3.3)	8.5 (3.2)	.66[c]	3.29 (199)	.001	0.23	−1.85 (199)	.033	−0.3, −0.1
Star Cancellation Test total score (*n* = 200)	52.7 (2.3)	52.8 (2.1)	.62[c]	0.69 (199)	.489	0.05	−4.40 (199)	<.001	−0.1, 0.1
Star Cancellation Test duration per correct cancellation (*n* = 199)	0.8 (0.2)	0.9 (0.3)	.68[c]	2.44 (198)	.016	0.17	2.64 (198)	.004	−0.2, 0

Note. In each comparison, a negative value means that the scores of the ISC test were larger than those of the analog version. The two one-sided t-tests (TOST) test for equivalence between scores by addressing the composite null-hypothesis that the observed difference between scores either exceeds the lower or exceeds the upper boundary value of a user-specified equivalence range (Lakens et al., 2018; see also the main text for more information). We defined 'practical equivalence' by means of lower and upper equivalence bounds of -0.3 *SD* and +0.3 *SD* of the scores of the analog test versions. The Table shows the one-sided*t*-test with the least extreme *p*-value (where a significant result is evidence for equivalence) and the corresponding 90% CI of the difference between the mean scores.ICC = Intraclass correlation; ISC = IntelliSpace Cognition; TMT = Trail Making Test; TOST = Two-one-sided t-tests equivalence test; RAVLT = Rey Auditory Verbal Learning Test; ROCFT = Rey-Osterrieth Complex Figure Test.

[a]Poor.

[b]Fair.

[c]Good.

[d]Excellent (Cicchetti, 1994).

tests except the ROCFT and TMT. For Letter Fluency, RAVLT Delayed Recall and Star Cancellation, paired samples t-tests indicated significant differences between the modalities with small effect sizes ranging from $d = 0.15$ to $d = 0.31$. However, the equivalence tests were significant, indicating that the differences in test scores fell within the *a priori* equivalence bounds of 0.3 *SD*s and can therefore be considered practically equivalent. We found non-equivalence for both TMT trials, the ROCFT Copy and ROCFT Immediate Recall. People tended to spend longer on the ISC version of the TMT compared to the analog version, constituting medium effect sizes for both trials. For ROCFT Copy, people tended to score higher on the ISC version compared to the analog version. The reverse was true for the ROCFT Immediate recall trial, where people tended to score lower on the ISC version compared to the analog version. For both ROCFT scores, the difference constituted a small effect size.

We conducted additional exploratory analyses to better understand the relative influence of the scoring method (i.e., automated vs. manual) separately from the administration method (i.e., digital vs. analog) on modality differences for the ROCFT. A subset of ROCFT drawings produced on the ISC platform ($n = 103$ for both the Copy and Immediate Recall trial) were manually scored by three independent neuropsychologists (each scoring a subset of the drawings). Thus, any effects that the ROCFT algorithm might have on scoring were removed in this comparison. Similar to the main analysis described above, we tested for equivalence between the manual scores for the ISC drawings and analog drawings using the TOST procedure.

For the Copy trial, the modalities were found to be equivalent when both were scored by human raters, (TOST: $t(102) = 2.08$, $p = .020$). The scores given by human raters on ISC drawings ($M = 23.9$, $SD = 6.2$) did not significantly differ from scores given by human raters on analog drawings ($M = 24.8$, $SD = 6.7$), $t(102) = 1.55$, $p = .125$. Similarly, we found equivalence for the Immediate Recall trial when both modalities were scored by the same human raters (TOST: $t(102) = 3.02$, $p = .002$). The scores given by the human raters on ISC drawings ($M = 10.1$, $SD = 4.9$) did not significantly differ from scores given by human raters on analog drawings ($M = 10.6$, $SD = 5.7$). These results suggest that scoring differences between the ISC version of the ROCFT and the analog version were the result of differences between the human raters and the algorithm, and not of differences in test performance.

Cross-modal measurement invariance

Multicollinearity statistics were acceptable for all measures (see Table S1 and Table S2 for correlations between all measures and Table S3 for VIF and tolerance values). The three RAVLT trials (Learning trial, Immediate Recall trial, and Delayed Recall trial) showed higher VIF and lower tolerance values than the other tests, with VIF values ranging between 3.54 (for analog learning trials) and 4.31 (for ISC immediate recall trial) and tolerance values ranging between 0.23 (for ISC immediate recall trial) and 0.28 (for analog learning trials). These values lie within acceptable boundaries (following common cut-off values of <5 for VIF and <.20 for tolerance). A control analysis, leaving out RAVLT Immediate Recall (which showed the highest VIF values and lowest tolerance) from the measurement invariance analyses yielded nearly identical results.

Standardized loadings for the configural model are presented in Table 5. The model converged normally. The chi-squared test was statistically significant (χ^2 (62) = 213.88, p = .013), suggesting a deviation of the sample covariance matrix from the model-implied covariance matrix and therefore suboptimal model fit. In contrast, the other fit statistics all reflected good model fit: RMSEA = .037, 90% CI [.018, .052], CFI = .981, SRMR = .048. Combined, this provided sufficient evidence for an adequate fit of the configural model. Thus, both the ISC and analog versions of the tests adhered to the same basic factor structure. See the Supplemental materials for the correlations between outcome measures.

The results of subsequent measurement invariance steps are summarized in Table 6. The model testing metric invariance in terms of equal loadings was supported. This demonstrates that the test scores contributed in the same way to the underlying cognitive domains across modalities. These results supported the hypothesis that the ISC test versions measured the same cognitive domains as their analog counterparts.

Constraining the test intercepts to be equal across modalities resulted in a significantly worse model fit. Thus, scalar invariance was not supported. In order to identify the tests driving the scalar non-invariance, we iteratively released intercept constraints based on modification indices. We did so one outcome measure at a time, each time comparing the updated partial scalar invariance model to the metric invariance model until the change in chi-squared was no longer statistically significant (see Table 7). Scalar invariance was most strongly driven by the intercept constraints of TMT A, TMT B and ROCFT Immediate Recall, with iterative modification indices ranging from 43.86 to 27.95. After releasing the intercept constraints of these tests, model fit was still significantly worse compared to the metric invariance model. Partial scalar invariance

Table 5. Standardized factor loadings based on analog and IntelliSpace cognition test versions.

	Analog			ISC		
	Estimate	*SE*	Z (*p*)	Estimate	*SE*	Z (*p*)
Executive functioning/speed						
TMT A duration	0.64	0.06	11.63*	0.61	0.06	10.45*
TMT B duration	0.76	0.05	16.19*	0.65	0.06	11.14*
Star cancellation duration per correct cancellation	0.49	0.06	7.78*	0.49	0.07	7.61*
Letter fluency total score	0.47	0.07	7.27*	0.53	0.06	8.60*
Working memory						
Digit Span Forward total score	0.72	0.07	10.87*	0.64	0.06	10.71*
Digit Span Backward total score	0.70	0.07	10.00*	0.75	0.06	12.40*
Memory						
RAVLT Learning total score	0.88	0.04	22.96*	0.87	0.03	27.05*
RAVLT Immediate Recall total score	0.92	0.02	55.37*	0.92	0.02	56.43*
RAVLT Delayed Recall total score	0.92	0.02	57.74*	0.90	0.02	48.99*
Visual spatial processing						
ROCFT Copy total score	0.73	0.07	9.94*	0.70	0.06	12.04*
ROCFT Immediate Recall total score	0.77	0.07	11.34*	0.73	0.07	13.41*

Note. For ease of interpretation, all test scores were coded so that lower scores reflected worse performance. ISC = IntelliSpace Cognition; TMT = Trail Making Test; RAVLT = Rey Auditory Verbal Learning Test; ROCFT = Rey-Osterrieth Complex Figure Test.

n = 189.

*p < .001.

Table 6. Measurement invariance of cognitive domains and test scores across analog and intelli Space cognition test modalities.

Model	χ^2 (df), p	RMSEA (90% CI)	CFI	SRMR	$\Delta\chi^2$	p
1. Configural invariance	213.88 (170), .013	.037 [.018, .052]	.981	.048		
2. Metric invariance: Equal factor loadings	228.97 (181), .009	.037 [.020, .052]	.979	.055	15.08	.179
3. Scalar invariance: Equal indicator intercepts	399.46 (192), <.001	.076 [.065, .086]	.901	.073	170.49	<.001

Note. CFI = Comparative Fit Index; CI = Confidence interval; SRMR = Standardized Root Mean Square Residual; RMSEA = Root Mean Square Error of Approximation.
$n = 189$.

Table 7. Partial scalar invariance analysis relative to the metric invariance model.

Test intercept	Modification index	$\Delta\chi^2$
TMT A duration	43.86***	116.97***
TMT B duration	44.05***	65.17***
ROCFT Immediate Recall total score	27.95***	32.69***
ROCFT Copy total score	6.99**	25.42***
RAVLT Delayed Recall total score	5.97*	19.32**
Digit Span Forward total score	5.24*	13.83*
Star Cancellation Test duration per correct cancellation	4.55*	9.17

Note. TMT = Trail Making Test; RAVLT = Rey Auditory Verbal Learning Test; ROCFT = Rey-Osterrieth Complex Figure Test.
$n = 189$.
*p < .05. **p < .01. ***p < .001.

was achieved only after additionally releasing the intercept constraints on ROCFT Copy, RAVLT Delayed Recall, Digit Span Forward, and Star Cancellation Test, although the modification indices were noticeable smaller (ranging from 6.99 to 4.55). Thus, only Digit Span Forward, RAVLT Learning, RAVLT Immediate Recall, and Letter Fluency Test showed scalar invariance across modalities.

Effects of demographics and experience with Digital devices

Table 8 compares the effects of age, years of education, sex, experience with digital devices, and modality order across modalities for each test. The residuals for TMT A duration, TMT B duration, and Star Cancellation Test duration per correct cancellation were found to be non-normally distributed. Therefore, we log-transformed the scores of these tests before submitting them to the regression analysis.

Overall, the strongest effects of demographic characteristics on performance were found for the RAVLT trials, showing comparable effects between modalities. In addition, the order in which the batteries were administered had a strong effect on all RAVLT trials. The pattern generally suggested a learning effect that was similar for both modalities: People tended to score lower on the RAVLT trials during their first visit than during their second visit, regardless of whether their first exposure was to the digital or analog version.

Age had a noticeably stronger negative effect on the ISC version of TMT A compared to the analog version, with older adults performing comparatively worse on the ISC version compared to the analog version. The effect of years of education

Table 8. Standardized regression coefficients and standard errors for the effect of demographics and version order on test scores across test modalities.

Test	Modality	Age	Years of Education	Sex[a]	Experience with digital devices	Modality order[b]	Adjusted R^2
TMT A duration	ISC	−0.24* (0.08)	0.02 (0.07)	−0.10 (0.07)	0.14 (0.08)	−0.13 (0.07)	.11
	Analog	−0.13 (0.08)	0.10 (0.08)	−0.12 (0.08)	0.11 (0.08)	0.10 (0.08)	.04
TMT B duration	ISC	−0.18 (0.08)	0.10 (0.07)	0.01 (0.07)	0.22* (0.08)	−0.10 (0.07)	.10
	Analog	−0.19* (0.08)	0.21* (0.07)	−0.04 (0.07)	0.15 (0.08)	0.09 (0.08)	.11
Digit Span Forward total score	ISC	−0.15 (0.08)	0.17 (0.07)	0.05 (0.07)	0.05 (0.08)	0.02 (0.07)	.03
	Analog	−0.15 (0.07)	0.22 (0.07)	−0.01 (0.07)	−0.02 (0.08)	0.18 (0.07)	.07
Digit Span Backward total score	ISC	−0.19 (0.07)	0.18 (0.07)	0.06 (0.07)	0.00 (0.07)	−0.06 (0.07)	.06
	Analog	−0.03 (0.08)	0.16 (0.07)	−0.05 (0.07)	0.04 (0.08)	0.03 (0.07)	.00
Letter Fluency total score	ISC	−0.02 (0.08)	0.16 (0.07)	−0.02 (0.07)	0.05 (0.08)	0.03 (0.07)	.00
	Analog	0.04 (0.08)	0.21* (0.07)	−0.11 (0.07)	0.11 (0.08)	−0.06 (0.07)	.05
ROCFT Copy total score	ISC	0.08 (0.08)	0.11 (0.07)	0.00 (0.07)	0.10 (0.08)	−0.12 (0.07)	.01
	Analog	−0.02 (0.08)	0.23* (0.07)	0.00 (0.07)	0.07 (0.08)	−0.13 (0.07)	.06
ROCFT Immediate Recall total score	ISC	−0.13 (0.07)	0.20* (0.07)	0.17* (0.07)	0.01 (0.07)	−0.20* (0.07)	.12
	Analog	−0.07 (0.08)	0.11 (0.07)	0.16 (0.07)	0.07 (0.08)	0.03 (0.07)	.04
RAVLT Learning total score	ISC	−0.25*** (0.07)	0.22** (0.06)	−0.29*** (0.06)	−0.08 (0.07)	−0.28*** (0.06)	.25
	Analog	−0.15* (0.07)	0.16* (0.07)	−0.34*** (0.07)	0.03 (0.07)	0.26*** (0.07)	.19
RAVLT Immediate Recall total score	ISC	−0.12 (0.07)	0.23** (0.07)	−0.21** (0.07)	0.07 (0.07)	−0.25*** (0.07)	.17
	Analog	−0.14 (0.07)	0.17* (0.07)	−0.30*** (0.07)	0.04 (0.07)	0.26*** (0.07)	.17
RAVLT Delayed Recall total score	ISC	−0.11 (0.07)	0.23** (0.07)	−0.25** (0.07)	0.03 (0.07)	−0.18* (0.07)	.13
	Analog	−0.11 (0.07)	0.20** (0.07)	−0.27** (0.07)	−0.01 (0.07)	0.20** (0.07)	.12
Star Cancellation Test duration per correct cancellation	ISC	−0.27** (0.08)	0.06 (0.07)	−0.12 (0.07)	0.08 (0.08)	−0.11 (0.07)	.10
	Analog	−0.32*** (0.08)	0.04 (0.07)	−0.09 (0.07)	0.12 (0.08)	0.08 (0.07)	.13

Note. For ease of interpretation, all test scores were coded so that lower scores reflected worse performance. *P*-values were corrected per test using the Benjamini-Hochberg procedure.
ISC = IntelliSpace Cognition; TMT = Trail Making Test; RAVLT = Rey Auditory Verbal Learning Test; ROCFT = Rey-Osterrieth Complex Figure Test.
$n = 184$ across all regression models.
[a]Dummy-coded using female as the reference category.
[b]Dummy-coded using analog – ISC as the reference category.
*$p < .05$. **$p < .01$. ***$p < .001$.

diverged between modalities on TMT B duration, and on both ROCFT Copy and Immediate Recall total scores, with more years of education being associated with better performance. The analog version of TMT B duration was strongly predicted by years of education, while the ISC version was not. Years of education predicted performance on the analog version of the ROCFT copy score but not in the ISC version, while the opposite was true for ROCFT Immediate Recall total score. For the remainder of the tests, demographic effects on test performance were comparable between the modalities. On a scale of 0 to 12, participants on average rated their experience with digital devices with 7 ($SD = 2.6$, range = 0-12), indicating moderate experience. Self-reported experience was correlated moderately with age ($r = -.31$, $t(366) = 6.50$, $p < .001$), and was not significantly correlated with years of education ($r = .08$, $t(366) = 1.52$, $p = .129$). Participant's self-reported experience with digital devices was only correlated with performance on TMT B duration, with people who indicated more experience being faster on the test. This effect came in the place of the age and education effects that were found for the analog version of TMT B.

Discussion

Overall, we found fair to excellent agreement between the ISC and analog test battery, with the majority of ISC tests showing equivalence to their analog counterparts. These findings are in line with those found for other digital cognitive test batteries (e.g., Björngrim et al., 2019; Weintraub et al., 2013). A few differences were found which were mostly limited to drawing tests. People tended to be slower on the ISC version of the TMT and age had a larger effect on the ISC version of TMT A. As for the ROCFT, people generally scored higher on the ISC copy trial than on the analog version of the test, but lower on the ISC immediate recall trial. Follow-up analyses showed that this difference disappeared when ISC drawings were rated by humans instead of the algorithm, suggesting that the difference is not related to differences between modalities (e.g., drawing on a tablet versus on paper) but rather by differences between algorithm scores and human scores.

Arguably the largest appeal of digital neuropsychological testing is the possibility of using automated scoring algorithms to assess performance. The current findings suggest that the performance of the scoring algorithms perform close to or on the level of human raters. First, interrater reliabilities between human raters and algorithms on the same input were reliably above .90, indicating high relative agreement. Second, the scores produced by speech recognition algorithms on the ISC tests showed high absolute agreement with the scores produced by human raters on the analog tests. For the ROCFT, high relative agreement was found with human raters, although scores produced by the algorithm were not equivalent to those of human raters in an absolute sense. This indicates that although the algorithm "ranked" participants in more-or-less the same way, the exact scores tended to be systematically different. For both drawing and speech-recognition algorithms it should be noted that the current findings only speak to their ability to approach human raters in terms of scoring, not to the extent to which their scores are based on the exact same components. For example, it is possible that the speech-recognition algorithm involved in Letter Fluency occasionally transcribes a wrong word (e.g., "for" instead of "four) that happens to be

correct under the scoring rules, or that the ROCFT algorithm structurally tends to miss one segment and mistakenly recognizes another segment. Future research should focus more in-depth on the exact process by which the algorithms arrive at certain scores, which could guide additional improvements to ultimately match – or even exceed – human raters.

Importantly, despite finding scalar non-invariance for most tests and differential effects of demographic characteristics, we generally established evidence of construct validity for the ISC battery through a measurement invariance analysis. Even though most tests showed differences in their mean intercepts, which relates to mean differences between test versions discussed above, the digital versions measure the same cognitive domains as their analog counterparts. This finding is important in a clinical context. While raw performance scores should not be directly compared across test versions, it indicates that clinicians do not have to adjust the way they interpret performance on a relative scale. Thus, the results underline the importance of separate normative data for digital tests, even if they are directly derived from existing analog tests (Bauer et al., 2012, American Educational Research Association et al., 2014).

A moderately strong relationship was found between age and experience with digital devices, with older people tending to use digital devices less frequently in their everyday life. However, the data showed that the average reported experience was relatively high, with only a handful of people indicating having (almost) no experience with digital devices. This is in line with the general trend in the population showing that older adults are increasingly familiar with digital technology (File & Ryan, 2014; Ryan, 2018). Importantly, TMT B was the only test in the ISC battery where performance was influenced by people's self-reported experience with digital devices. The effect of experience might have been somewhat suppressed by the high age of the sample, although the fact that even in this relatively older sample the self-reported experience was high suggests that potential effects in a broader age range might not be as important as reported elsewhere (e.g., Hauk et al., 2018; Lee Meeuw Kjoe et al., 2021; Rabipour & Davidson, 2020; Wild et al., 2012). It is possible that the level of digital experience observed in the current sample was somehow biased and non-representative, although a substantial bias seems unlikely given the strict sampling plan based on education and ethnicity.

The current study has some limitations that should be considered. First, as noted in the previous paragraph, the age of the sample ranges between 50 and 80 years. Therefore, the results cannot be extrapolated to individuals that fall outside this age range. A quantification of age effects (as well as other demographic effects) across a broader age range is an important step for future research. Second, the sample only includes healthy participants. Converting cognitive tests to a digital format can have a different impact for people with cognitive impairments compared to healthy people. For example, differences on drawing tests might be magnified for patients with impairments in motor or perceptual domains (Germine et al., 2019). Such differences could potentially have an impact on the construct validity of the tests, as well as their clinical sensitivity for specific cognitive impairments. An important future extension of the current study, therefore, will be to administer the ISC battery to patients with a variety of cognitive impairments. Third, while we focused on the

effects of several demographic characteristics across a representative sample, other factors such as ethnicity, acculturation or quality of education were not considered.

To conclude, the current study provides evidence for criterion and construct validity of the ISC cognitive test battery. Most ISC tests yielded scores that were equivalent to their analog counterparts, both in absolute and relative terms. Some differences were found for the drawing tests TMT and ROCFT, which both showed differences in mean scores and influences from demographic characteristics. However, additional analyses indicated that the differences found for ROCFT were caused by differences between algorithm and human rater scores, and not by structural differences between modalities. Importantly, metric invariance was established for the entire test battery, meaning that the ISC tests measure the same cognitive constructs as their analog counterparts. Finally, people's experience with digital devices was found to only affect TMT B performance, suggesting that the battery is suitable for use in the older population. These findings offer an important step in validating the ISC cognitive test battery for clinical use.

Acknowledgments

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

Stefan Vermeent, Mandy Spaltman, Gijs van Elswijk, and Ben Schmand were employed by Philips. Justin B. Miller received consultation fees from Philips.

ORCID

Stefan Vermeent http://orcid.org/0000-0002-9595-5373
Justin B. Miller http://orcid.org/0000-0002-4439-6604

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). Standards for Educational and Psychological Testing. American Educational Research Association.

Arrieux, 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

Bailey, S. K. T., Neigel, A. R., Dhanani, L. Y., & Sims, V. K. (2018). Establishing measurement equivalence across computer- and paper-based tests of spatial cognition. *Human Factors*, *60*(3), 340–350. https://doi.org/10.1177/0018720817747731

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. *Archives of Clinical Neuropsychology*, *27*(3), 362–373. https://doi.org/10.1093/arclin/acs027

Benton, A. I., & Hamsher, K. (1989). *Multilingual aphasia examination*. AJA Associates.

Björngrim, S., van den Hurk, W., Betancort, M., Machado, A., & Lindau, M. (2019). Comparing traditional and digitized cognitive tests used in standard clinical evaluation - a study of the digital application minnemera. *Frontiers in Psychology*, *10*, 2327. https://doi.org/10.3389/fpsyg.2019.02327

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cole, W. R., Arrieux, J. P., Ivins, B. J., Schwab, K. A., & Qashu, F. M. (2018). A comparison of four computerized neurocognitive assessment tools to a traditional neuropsychological test battery in service members with and without mild traumatic brain injury. *Archives of Clinical Neuropsychology : The Official Journal of the National Academy of Neuropsychologists*, *33*(1), 102–119. https://doi.org/10.1093/arclin/acx036

Cole, D., & Maxwell, S. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology*, *112*(4), 558–577.https://doi.org/10.1037/0021-843X.112.4.558

Core Team, R. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Dadey, N., Lyons, S., & DePascale, C. (2018). The comparability of scores from different digital devices: A literature review and synthesis with recommendations for practice. *Applied Measurement in Education*, *31*(1), 30–50. https://doi.org/10.1080/08957347.2017.1391262

Daniel, M. H., Wahlstrom, D., & Zhang, O. (2014). *Equivalence of Q-interactive™ and paper administrations of cognitive tasks: WISC®-V*. Q-Interactive Technical Report 8.

File, T., & Ryan, C. (2014). *Computer and internet use in the United States: 2013*. US Census Bureau.

Folstein, M. F., Folstein, S. E., White, T., & Messer, M. A. (2010). *MMSE-2: Mini-mental state examination* (2nd ed.).PsychologicalAssessmentResources.

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

Grinspoon, L., Passell, E., Scheuer, L., Ressler, K., & Germine, L. (2019). F147. Going digital? Understanding the impact of technical variability on neurocognitive assessment. *Biological Psychiatry*, *85*(10), S270. https://doi.org/10.1016/j.biopsych.2019.03.684

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

Hauk, N., Hüffmeier, J., & Krumm, S. (2018). Ready to be a silver surfer? A meta-analysis on the relationship between chronological age and technology acceptance. *Computers in Human Behavior*, *84*, 304–319. https://doi.org/10.1016/j.chb.2018.01.020

Hülür, G., & MacDonald, B. (2020). Rethinking social relationships in old age: Digitalization and the social lives of older adults. *American Psychologist*, *75*(4), 554–566. https://doi.org/10.37/amp0000604

Kessels, R. P. C. (2019). Improving precision in neuropsychological assessment: Bridging the gap between classic paper-and-pencil tests and paradigms from cognitive neuroscience. *The Clinical Neuropsychologist*, *33*(2), 357–368. https://doi.org/10.1080/13854046.2018.1518489

Koo, B. M., & Vizer, L. M. (2019). Mobile technology for cognitive assessment of older adults: A scoping review. *Innovation in Aging*, *3*(1), 2–14. https://doi.org/10.1093/geroni/igy038

Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Lee Meeuw Kjoe, P. R., Agelink van Rentergem, J. A., Vermeulen, I. E., & Schagen, S. B. (2021). How to correct for computer experience in online cognitive testing? *Assessment*, *28*(5), 1247–1255. https://doi.org/10.1177/1073191120911098

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological Assessment*. (5th ed.). Oxford University Press.

Mackinnon, S. P., Curtis, R., & O'Connor, R. M. (2021). *A tutorial in longitudinal measurement invariance and cross-lagged panel models using Lavaan*. PsyArXiv.

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. *The International Journal of Psychiatry in Medicine*, *24*(3), 229–244. https://doi.org/10.2190/5A0F-936P-VG8N-0F5R

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: The Official Journal of the National Academy of Neuropsychologists*, *32*(5), 541–554. https://doi.org/10.1093/arclin/acx050

Morrison, R. L., Pei, H., Novak, G., Kaufer, D. I., Welsh-Bohmer, K. A., Ruhmel, S., & Narayan, V. A. (2018). A computerized, self-administered test of verbal episodic memory in elderly patients with mild cognitive impairment and healthy participants: A randomized, crossover, validation study. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *10*(1), 647–656. https://doi.org/10.1016/j.dadm.2018.08.010

Passell, E., Rutter, L. A., Kim, H., Scheuer, L., Martini, P., Grinspoon, L., & Germine, L. (2020). *Cognitive test scores vary based on choice of personal digital device*. PsyArXiv. https://doi.org/10.31234/osf.io/wbmdj

Rabipour, S., & Davidson, P. S. R. (2020). Using digital technology for cognitive assessment and enhancement in older adults. In M. N. Potenza, K. A. Faust, & D. Faust (Eds.), *The Oxford handbook of digital technologies and mental health* (pp. 358–372). Oxford University Press.

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

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Ryan, C. (2018). *Computer and internet use in the United States: 2016*. US Census Bureau.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*, 23–74.

Schlegel, R. E., & Gilliland, K. (2007). Development and quality assurance of computer-based assessment batteries. *Archives of Clinical Neuropsychology*, *22S*, S49–S61. https://doi.org/10.1016/j.acn.2006.10.005

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., & den Dungen, M. (2008). Letterfluency: Psychometrische eigenschappen en Nederlandsenormen [LetterFluency: Psychometriccharacteristics and Dutch norms]. *Tijdschrift Voor Gerontologie en Geriatrie*, *39*(2), 64–74. https://doi.org/10.1007/BF03078128

Schmidt, M. (1996). *Rey auditory verbal learning test: A handbook*. Western Psychological Services.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680. https://doi.org/10.1007/BF01068419. PMID: 3450848

Spreij, L. A., Gosselt, I. K., Visser-Meily, J. M. A., & Nijboer, T. C. W. (2020). Digital neuropsychological assessment: Feasibility and applicability in patients with acquired brain injury. *Journal of Clinical and Experimental Neuropsychology*, *42*(8), 781–793. https://doi.org/10.1080/13803395.2020.1808595

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary (third edition)*. Oxford University Press.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Vermeent, S., Dotsch, R., Schmand, B., Klaming, L., Miller, J. B., & van Elswijk, G. (2020). Evidence of validity for a newly developed digital cognitive test battery. *Frontiers in Psychology*, *11*, 1–11. https://doi.org/10.3389/fpsyg.2020.00770

Weintraub, S., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Zelazo, P. D., Bauer, P. J., Carlozzi, N. E., Slotkin, PJ., 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, Supplement 3), S54–S64. https://doi.org/10.1212/WNL.0b013e3182872ded.

Wild, K., Howieson, D., Webbe, F., Seelye, A., & Kaye, J. (2008). The status of computerized cognitive testing in aging: A systematic review. *Alzheimer's & Dementia*. *Alzheimer's & Dementia*, *4*(6), 428–437. https://doi.org/10.1016/j.jalz.2008.07.003

Wild, K. V., Mattek, N., Maxwell, S. A., Dodge, H. H., Jimison, H. B., & Kaye, J. A. (2012). Computer related self-efficacy and anxiety in older adults with and without mild cognitive impairment. *Alzheimer's & Dementia*, *8*(6), 544–552. https://doi.org/10.1016/j.jalz.2011.12.008

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 & Other Dementiasr*, *30*(1), 13–28. https://doi.org/10.1177/1533317514522852