Reliability and validity of a self-administered tool for online neuropsychological testing: The Amsterdam Cognition Scan

Feenstra, Heleen E.M.; Murre, Jaap M.J.; Vermeulen, Ivar E.; Kieffer, Jacobien M.; Schagen, Sanne B.

published in
Journal of Clinical and Experimental Neuropsychology
2018

DOI (link to publisher)
10.1080/13803395.2017.1339017

document version
Publisher's PDF, also known as Version of record

Routledge
Taylor & Francis Group

Journal of Clinical and Experimental Neuropsychology

ISSN: 1380-3395 (Print) 1744-411X (Online) Journal homepage: https://www.tandfonline.com/loi/ncen20

Reliability and validity of a self-administered tool for online neuropsychological testing: The Amsterdam Cognition Scan

Heleen E. M. Feenstra, Jaap M. J. Murre, Ivar E. Vermeulen, Jacobien M. Kieffer & Sanne B. Schagen

To cite this article: Heleen E. M. Feenstra, Jaap M. J. Murre, Ivar E. Vermeulen, Jacobien M. Kieffer & Sanne B. Schagen (2018) Reliability and validity of a self-administered tool for online neuropsychological testing: The Amsterdam Cognition Scan, Journal of Clinical and Experimental Neuropsychology, 40:3, 253-273, DOI: 10.1080/13803395.2017.1339017

To link to this article: https://doi.org/10.1080/13803395.2017.1339017

JOURNAL OF CLINICAL AND EXPERIMENTAL NEUROPSYCHOLOGY, 2018
VOL. 40, NO. 3, 253–273
https://doi.org/10.1080/13803395.2017.1339017

Routledge
Taylor & Francis Group

Reliability and validity of a self-administered tool for online neuropsychological testing: The Amsterdam Cognition Scan

Heleen E. M. Feenstra[a], Jaap M. J. Murre[b], Ivar E. Vermeulen[c], Jacobien M. Kieffer[a] and Sanne B. Schagen[a,b]

[a]Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute, Amsterdam, The Netherlands; [b]Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands; [c]Department of Communication Science, VU University Amsterdam, Amsterdam, The Netherlands

ABSTRACT

Introduction: To facilitate large-scale assessment of a variety of cognitive abilities in clinical studies, we developed a self-administered online neuropsychological test battery: the Amsterdam Cognition Scan (ACS). The current studies evaluate in a group of adult cancer patients: test–retest reliability of the ACS and the influence of test setting (home or hospital), and the relationship between our online and a traditional test battery (concurrent validity). **Method**: Test–retest reliability was studied in 96 cancer patients (57 female; M_{age} = 51.8 years) who completed the ACS twice. Intraclass correlation coefficients (ICCs) were used to assess consistency over time. The test setting was counterbalanced between home and hospital; influence on test performance was assessed by repeated measures analyses of variance. Concurrent validity was studied in 201 cancer patients (112 female; M_{age} = 53.5 years) who completed both the online and an equivalent traditional neuropsychological test battery. Spearman or Pearson correlations were used to assess consistency between online and traditional tests. **Results**: ICCs of the online tests ranged from .29 to .76, with an ICC of .78 for the ACS total score. These correlations are generally comparable with the test–retest correlations of the traditional tests as reported in the literature. Correlating online and traditional test scores, we observed medium to large concurrent validity (r/ρ = .42 to .70; total score r = .78), except for a visuospatial memory test (ρ = .36). Correlations were affected—as expected—by design differences between online tests and their offline counterparts. **Conclusions**: Although development and optimization of the ACS is an ongoing process, and reliability can be optimized for several tests, our results indicate that it is a highly usable tool to obtain (online) measures of various cognitive abilities. The ACS is expected to facilitate efficient gathering of data on cognitive functioning in the near future.

ARTICLE HISTORY
Received 22 August 2016
Accepted 28 May 2017

KEYWORDS
Amsterdam Cognition Scan; computer skills; concurrent validity; intraclass correlation coefficient; neuropsychological test battery; oncology; test development; test–retest reliability; unmonitored setting

Neuropsychological testing is the main method used to measure cognitive functioning (Gates & Kochan, 2015; Lezak, Howieson, Loring, Hannay, & Fischer, 2004), but the time-consuming and labor-intensive nature of traditional, supervised neuropsychological assessments limits its application. For cognitive studies with highly diverse study samples, for example, there is a need for more efficient ways to collect neuropsychological data. Similarly, when time frames for testing are short, efficiency is also warranted. In these cases traditional face-to-face assessment methods may not be feasible, as they severely limit both the scale and speed of data collection (Caine, Mehta, Laack, & Gondi, 2012; Van de Weijer-Bergsma, Kroesbergen, Prast, & Van Luit, 2015). Clearly, there is a need for efficient neuropsychological measurement tools that allow for standardized, large-scale data collection for cognitive research.

In the context of this need, computerized testing has been used more and more to assess cognitive functioning during the past decade, both for research and for clinical practice (Bauer et al., 2012; Iverson, Brooks, Ashton, Johnson, & Gualtieri, 2009). The main advantages of computerized testing are standardization (Barak & English, 2002; Bilder, 2011; Parsey & Schmitter-Edgecombe, 2013; Reips, 2002a), additional observations (e.g., reaction times) (Barak & English, 2002; Bauer et al., 2012; Naglieri et al., 2004; Schatz & Browndyke, 2002), and reduced costs for test administration and scoring (Bauer et al., 2012). However, computerized assessments often require supervision and are typically bound to specific test locations. As computers and internet connections are currently available in most western households (Internet World Stats, 2017), *online* neuropsychological testing potentially offers an efficient

CONTACT Sanne B. Schagen s.schagen@nki.nl The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands
The research was conducted at the The Netherlands Cancer Institute, Amsterdam, The Netherlands.
The supplemental material for this article can be accessed here.

and cost-effective way to significantly increase sample sizes for cognitive studies.

Online testing, as a subtype of computerized testing, has the added advantages of higher accessibility and efficiency. There are few restrictions on the timing and the location of an online assessment (Barak & Buchanan, 2003; Birnbaum, 2004; Caine et al., 2012; Naglieri et al., 2004; Reips, 2002b), especially when such assessments are unmonitored (i.e., without a test leader present). Participants can take the test from home, on their own computer, and without having to install additional software. As it is unnecessary to travel to a specific test location, people in remote locations, with mobility problems, or with busy lifestyles may be more likely to take part in research studies (Birnbaum, 2004; Buchanan & Smith, 1999; Caine et al., 2012; Germine et al., 2012; Naglieri et al., 2004; Reips, 2002b), which in turn may reduce participation bias. The self-administration of the tests also reduces costs for labor, equipment, and travel (Bauer et al., 2012; Buchanan & Smith, 1999; Caine et al., 2012; Naglieri et al., 2004; Reips, 2002a; Van de Weijer-Bergsma et al., 2015). Furthermore, during unmonitored testing, people may be more willing to answer personal questions candidly (Gunter, Nicholas, Huntington, & Williams, 2002; Reips, 2002a, 2002b, 2002c).

Online testing tools do have possible limitations as well. The fact that (as self-administered tests) they take place in an unmonitored setting generally makes the tests more sensitive to distraction and misinterpretation in participants. Moreover, online tests are —like all computerized tests—sensitive to variability in participants' computer skills and computer configurations (Bauer et al., 2012; Parsey & Schmitter-Edgecombe, 2013). To minimize such contextual influences, online neuropsychological tests should meet stringent criteria. First, tests should be designed for the unmonitored setting, focusing on: clear instructions, on optimizing participants' concentration and motivation (Cromer et al., 2015), and on compatibility with a large variety of operating systems, computer hardware, and internet browsers (Reips, 2002a). Second, reliability and validity of the tests across different contexts should be properly studied and reported (Gates & Kochan, 2015; Naglieri et al., 2004). Third, online-specific normative data should be collected and accessible (Bauer et al., 2012). The online test batteries that are currently available often still require on-site supervision by a test leader or lack proper psychometric evaluation and norm data. For example, ImPACT, probably the most widely used online neuropsychological test battery in the world (Tsushima et al., 2013), is not suitable for unmonitored assessments (ImPACT Applications. Inc., n.d.), whereas WebNeuro (Silverstein et al., 2007), another established online neuropsychological test battery, is suitable for unmonitored assessments, but lacks online-based normative data. Furthermore, all other currently available tests that are suitable for self-administered assessments are limited to detecting age-related cognitive decline and/or are developed for short cognitive screening. Therefore, we have developed a self-administered online test battery covering a broad range of cognitive functions and suitable for a wide age range.

Because in oncological research, as in many other research fields (e.g., cognitive aging), general cognitive functioning needs to be studied in large patient samples to answer important questions on incidence rates, risk factors, and long-term trajectory of the development of cognitive impairment (Schagen et al., 2014), and since it is challenging and very costly to obtain large (clinical) data sets using traditional neuropsychological tests alone, we developed the Amsterdam Cognition Scan (ACS): a new online neuropsychological test battery that can be completed without supervision. Currently, a Dutch and an English version are available. The battery is based on seven well-established traditional neuropsychological tests, which were chosen to detect dysfunction throughout the spectrum of cognition, including the domains attention, information processing speed, learning and memory, executive functioning, and psychomotor speed.

As mentioned above, test development requires extensive psychometric evaluation. Reliability and validity of the new test need to be assessed, and normative standards should be determined. In order to interpret test results administered from both the home and the hospital setting, it is also important to evaluate the influence of test setting on performance. Home assessments are less structured than hospital assessments, and variability in technology (e.g., operating system, external hardware, and browser) and distractions during the assessment could influence performance and test reliability. Therefore, the primary aims of the current studies are to assess test–retest reliability (consistency of test scores over time) and the effect of test setting (Experiment 1), and to assess concurrent validity (comparability to a previous validated measure; Experiment 2) of the Dutch version of the ACS. Both studies use patients from an oncological setting. Reference data of 249 healthy controls have been collected (Feenstra, Vermeulen, Murre, & Schagen, in press) and will be collected continuously to determine normative standards.

Method

Participants and procedure

All participants were recruited via the Netherlands Cancer Institute (October 2013 to November 2014). Patients treated for noncentral nervous system cancer between 2010 and 2013 were screened based on information from electronic patient files and were subsequently invited by mail. Exclusion criteria were: (a) tumor or metastases in the central nervous system; (b) distant metastases; (c) disease progression; and (d) psychiatric/neurologic symptoms hampering test completion. Inclusion criteria were: (a) age between 18 and 76 years; and (b) treatment with chemotherapy, radiotherapy, hormonal therapy, or immunotherapy. Eligible patients were randomized to participate either in Experiment 1 (test–retest reliability; one third of the patients), or in Experiment 2 (concurrent validity; two thirds of the patients). Patients were stratified by age (18–40; 41–59; 60–76 years), gender, and tumor type (one third breast; one third testis/prostate; one third other). One week after sending out invitation letters, patients were contacted by telephone. Patients who were interested in participation were asked a few questions to verify final eligibility. Participants were required to have basic computer skills and sufficient proficiency in the Dutch language (for both studies), and participants had to have access to a computer with internet connection (Experiment 1). If eligible, patients were asked to provide additional demographic and disease-related information.

The review board of the Netherlands Cancer Institute approved the study, which conformed with ethical guidelines for human experimentation stated in the Declaration of Helsinki. Written informed consent was obtained for all participants before the start of the assessments.

Experiment 1: Test–retest reliability and effect of test setting

Experiment 1 was designed to investigate: (a) test–retest reliability, and (b) the effect of the setting of the assessment on test performance. The latter was done by comparing patients' scores on the ACS from the hospital setting to their scores from the home setting. To facilitate a counterbalanced design, patients were randomly assigned either to complete the online test first at home and then at the hospital (HOME-HOSP), or to complete the online test first at the hospital and then at home (HOSP-HOME). The first and second assessments were planned to be 6 weeks apart. This interval was chosen in order to be long enough to resemble a clinical research setting while at the same time short enough to limit the chance of true changes in the patient population.

Experiment 2: Concurrent validity

Experiment 2 was designed to investigate how well the ACS's tests measure the constructs they aim to measure. We studied concurrent validity by comparing online test scores to scores on well-established traditional counterparts (de Vet, Terwee, Mokkink, & Knol, 2011). Participants were scheduled to visit the hospital once to complete both the self-administered online test battery and a matching supervised traditional test battery. Randomization created two subgroups for counterbalancing. Subgroup 1 first completed the traditional assessment and then the online assessment (T-O), whereas Subgroup 2 first completed the online assessment and then the traditional assessment (O-T). The online and the traditional assessments took about 1 hour and 15 min to complete each. The entire session was scheduled to take 3 hours, including a 15-min break in between the two assessments.

Development and characteristics of the online portal

The ACS was developed by a professional programmer (with experience in programming neuropsychological tests) and the project team. The online neuropsychological tests were based on widely used, traditional neuropsychological tests. The goal was to develop online tests that measure the same cognitive constructs as the traditional tests do as well as possible. The portal was extensively tested on usability with input from several other neuropsychologists, a human–computer interaction expert, and focus groups of cancer patients of varying age, gender, and education level. Three rounds of pilot testing and revisions were completed in order to develop the current version of the ACS.

The resulting online test battery is self-administrative and developed for assessments in unmonitored settings, such as from home. First, test performance is influenced as little as possible by technical variance. Minimum requirements are a computer with a sound system and an internet connection (mobile devices are not compatible). The tests are programmed to run on all major internet browsers—no downloads are required—and can be run on all common operating systems. Client-side programming enables reliable data collection, which is independent of internet speed or bandwidth.

Second, all elements of the test battery are self-explanatory. Overall test design and instructions were developed with the aim to require little cognitive load and to optimize motivation. Instructions are presented

through an audiovisual mode (video), promoting "dual processing" that fosters better understanding and learning as cognitive overload is prevented, and individual capacity is optimized (Brünken & Plass, 2003; van Hooijdonk & Krahmer, 2008). The test session starts with a general instruction video. In this video, a woman explains what to expect of the tests (duration and general content) and how to best prepare for the assessment (see online supplementary material)—for example: "Make sure you cannot be disturbed by the telephone, that you are alone in the room, and that the radio and television are turned off." Besides a purely informative function, this video aims to create a trustworthy and serious setting and to minimize impeding psychological factors such as fatigue and anxiety. Before the start of the neuropsychological tests, keyboard typing skills, mouse click skills, and mouse drag skills are assessed to provide a measure of computer skills (see Appendix A for a description of the Type Skills, Click Skills, and Drag Skills tests).

Next, is the neuropsychological test battery itself. Every test starts with its own instruction video; these videos consist of screen captures of the tests and a voice-over providing further instructions (see online supplementary material for an example). To make sure that participants are ready to proceed with the actual test, the instruction videos are followed by a practice session with feedback. Tests either have an error-based stop-rule or a time limit.

During all elements of the ACS, participants can use the "contact button" at the bottom of the screen to report problems or peculiarities. Two animation videos on a popular psychological topic were incorporated in the test battery to serve as fixed, standardized breaks. Moreover, participants are encouraged to complete the entire battery in one session, but in case of an unavoidable interruption the ACS automatically resumes at the start of the last incomplete element when a participant resumes the test at a later point in time. All input is saved on a dedicated and protected server after completion of each test element.

Test administration

The hospital assessments (online and traditional) took place in a quiet test room, which was set up with either a desktop computer with a 19" screen or a laptop with a 17" screen, both using Microsoft Windows 7 and Google Chrome to run the ACS. Sound stimuli were played by an attached speaker set. A standard QWERTY keyboard and a two-button mouse were used to record response input. All online assessments were completed in an unmonitored setting. In the hospital, online assessments were briefly introduced by one of three trained research assistants; as soon as the participant indicated to be comfortable, the assistant left the room, letting the participant complete the test in an unmonitored manner. There was a telephone in the test room, which allowed participants to call the assistant if there were any problems or when they had finished the tests. All face-to-face assessments were conducted by the same research assistant as the one who introduced the online test using standardized instructions.

The home assessments (online; Experiment 1) were completed from a private setting on a personal computer with internet connection. Participants were instructed to find a quiet environment without distractions and could, in case of any technical problems, use the contact button or directly call a research assistant to request for help.

Materials

Traditional neuropsychological test battery

The traditional neuropsychological test battery consisted of established tests that require face-to-face assessment (see Table 1): (a) Trail Making Test (TMT) A and B (Reitan, 1958); (b) Dutch version of the Rey Auditory Verbal Learning Test, short form (Van den Burg, Saan, & Deelman, 1985); (c) FePsy Visual Reaction Time (Alpherts & Aldenkamp, 1995); (d) Tower of London, Drexel University (TOL-dx) (Culbertson & Zillmer, 2001); (e) Corsi Block-Tapping Test (Kessels, van Zandvoort, Postma, Kappelle, & Edward, 2010); (f) Grooved Pegboard (Kløve, 1963); (g) Wechsler Adult Intelligence Scale–Third Edition (WAIS–III) Digit Span, forward & backward (Wechsler, 1997); and (h) the Dutch letter combination DAT for letter fluency (Schmand, Groenink, & van den Dungen, 2008). In addition, the Dutch reading test for adults (Schmand, Bakker, Saan, & Louman, 1991) was conducted to estimate level of IQ.

Online neuropsychological test battery: The Amsterdam Cognition Scan

The online neuropsychological test battery consisted of seven neuropsychological tests on the domains of attention, information processing speed, working memory, verbal learning and memory, visuospatial memory, executive functioning, and psychomotor speed. Table 1 describes all test domains and main outcome measures of the online tests and their traditional equivalents. A detailed overview of all elements of the ACS is given in Appendix A.

Table 1. The online tests and their equivalent traditional tests.

	Online tests	Test domains	Main outcome measure	Traditional equivalent
1	Connect the Dots I; Connect the Dots II	Visuomotor tracking, planning, cognitive flexibility, divided attention	Completion time (I & II)	Trail Making Test A Trail Making Test B
2a	Wordlist Learning	Verbal learning	Total number of correct words (Trials 1 to 5)	15 Words test (Dutch version of Rey Auditory Verbal Learning Test)
2b	Wordlist Delayed Recall & Recognition	Retention of information: free recall and recognition	Total number of correct words; free recall and recognition	15 Words test
3	Reaction Speed	Information processing speed & attention	Mean reaction time	Visual Reaction Time (subtest FePsy)
4	Place the Beads	Planning, response inhibition, visuospatial memory	Total number of extra moves	Tower of London, Drexel University (TOL-dx)
5	Box Tapping	Visuospatial short term memory	Total number of correctly repeated sequences	Corsi Block-Tapping Test
6	Fill the Grid	Fine motor skills	Completion time	Grooved Pegboard
7	Digit Sequences I; Digit Sequences II	I: Attention II: Working memory	Total number of correctly repeated sequences (I & II)	WAIS–III Digit Span (forward & backward)

Questionnaires

Following completion of both the online and the traditional neuropsychological test battery, two questionnaires were presented to complete online or via paper and pencil, respectively: the Hospital Anxiety and Depression Scale (HADS; Zigmond & Snaith, 1983), assessing symptoms of depression and anxiety, and the Multidimensional Fatigue Inventory (MFI–20; Smets, Garssen, Bonke, & De Haes, 1995), assessing symptoms of fatigue. In addition, in the online mode, directly after finishing these questionnaires, participants were online debriefed on test conditions during the assessment (e.g., disruptions and technical issues) and on qualitative appreciation of several elements of the ACS. Participants also had the opportunity to leave additional comments and tips regarding the ACS.

Statistical analyses

Patient characteristics

Comparison of the baseline characteristics of the subgroups of Experiment 1 and Experiment 2 were performed using independent-sample *t* tests and chi-square tests.

Outliers

Outliers on neuropsychological test scores (traditional and online) were identified and excluded from data analyses in order to limit their influence on correlation coefficients (Devlin, Gnanadesikan, & Kettenring, 1975). For reaction time outcomes (TMT/Connect the Dots, Visual Reaction Time/Reaction Speed, and Grooved Pegboard/Fill the Grid), we used the median absolute distance (MAD) to detect outliers (Leys, Ley, Klein, Bernard, & Licata, 2013). To prevent age-based bias, MADs were calculated and applied separately for the age groups: ≤40; 41–59; and ≥60 years. For tests that rely on number of correct responses and for which zero-scores are more likely to reflect usability issues than floor performance (15 Words test/Wordlist Learning, Corsi Block-tapping/Box Tapping, and WAIS–III Digit Span/Digit Sequences), zero-scores were considered outliers and were excluded from analyses.

Composite scores

In addition to analyses on the individual tests, composite scores of the total neuropsychological test batteries were calculated as the mean of the (reversed) *z*-scores of all main online neuropsychological outcome measures ("online total score") and all main traditional neuropsychological outcome measures ("traditional total score").

Effect of test setting and test–retest reliability (Experiment 1)

To assess the effect of test setting (home or hospital), subgroup, and time on the neuropsychological test scores, we used a mixed-effects model for repeated measures with random intercept, a maximum likelihood solution, and variance component covariance structure. Interaction effects of "Subgroup × Time" were tested, to indicate possible carryover effects of setting over time. If no such effects were found, we analyzed the main effects "subgroup" and "time."

Test–retest reliability was assessed with intraclass correlation coefficients (ICCs) as main outcome measure, since, unlike other correlation measures, ICCs can detect both random and systematic error (e.g., practice effects) when calculating consistency between assessments (de Vet et al., 2011). In recent years, ICCs are increasingly used to examine test–retest reliability for cognitive testing (Darby et al., 2014; Dougherty et al., 2010; Hansen, Lehn, Evensmoen, & Håberg, 2016; Littleton, Register-Mihalik, & Guskiewicz, 2015;

Medalia, Lim, & Erlanger, 2005; Ruano et al., 2016; Schatz, 2010). ICCs represent the proportion of repeated measures variance that is attributable to true variance, where higher ICC values indicate less error variance and better test–retest reliability (Weir, 2005). ICC values depend on between-subject variability (ICCs are larger if performance differs more between participants) and on type of ICC (Weir, 2005). Consequently, interpretation of ICC values is not unambiguous. However, generally, as a criterion for acceptable test–retest reliability, ICC values of .60 or .70 and higher are recommended (Anastasi, 1998; Cole et al., 2013; de Vet et al., 2011; Ruano et al., 2016). In agreement with these recommendations, we used a criterion of .60 to indicate which tests have sufficient reliability levels and which tests do not. To control for systematic error originating from setting between Assessment 1 and Assessment 2, we calculated the ICC using the within-subject variance and variance estimates of the factors "setting," "time," and "error" from mixed-effect models for repeated measures: ICC = $var(patients)/[var(patients) + var(time) + var(setting) + var(error)]$ (Weir, 2005). To enable comparisons between the test–retest reliabilities of our newly developed tests and reported test–retest reliabilities in the literature, Pearson's r and Spearman's rho (depending on the distributions of scores on the particular measurements) were also calculated.

Concurrent validity (Experiment 2)

Concurrent validity of the online tests was assessed with Spearman and Pearson correlation coefficients (again depending on the distributions of scores on the particular measurements), measuring consistency between the online and traditional test scores. Perfect correlations were not expected since, in several tests, we did not only apply those adaptations necessary to change traditional tests into online tests, but also applied additional adaptations to optimize assessment of the measurement construct. Therefore, we interpret correlations of ≥.40 to indicate acceptable validity (Trustram Eve & de Jager, 2014). To further interpret validity results adequately, reliability of the online and traditional measures should be taken into account. Validity coefficients were expected not to exceed: $\sqrt{(reliability\ traditional\ test \times reliability\ new\ test)}$ (de Vet et al., 2011). This "ceiling consistency" was calculated, using for the traditional tests the highest Pearson's r as reported in the literature, and for the online tests the Spearman's rho or Pearson's r as found in Experiment 1. A correction for effect of assessment order and differences in practice effects between the O-T and the T-O subgroups was applied to all second assessment test scores by subtracting the difference in mean test scores between subgroup T-O and O-T from the individual test scores. This means that the online scores for subgroup T-O are:

$$score\_online\_T\text{-}O\_corrected = score\_online\_T\text{-}O - (mean\_score\_online\_T\text{-}O - mean\_score\_online\_O\text{-}T),$$

and the traditional scores for subgroup O-T are:

$$score\_traditional\_O\text{-}T\_corrected = score\_traditional\_O\text{-}T - (mean\_score\_traditional\_O\text{-}T - mean\_score\_traditional\_T\text{-}O).$$

Influencing factors on online neuropsychological test scores

Multiple linear regression analyses were used to assess the predictive values of traditional test scores, age, gender (0 = female, 1 = male), IQ, computer experience, and computer skills on online test scores.

To explore the moderating effect of computer skills on test performance (Bauer et al., 2012; Iverson et al., 2009), a composite score was constructed by calculating the mean of the z-scores of completion times (time at last response – time at first response) of the type skills, click skills, and drag skills from the first online assessment.

All statistical analyses were performed using IBM SPSS Statistics for Windows, Version 22.0 (Armonk, NY: IBM corp.). For correlational analyses, probabilities of $p < .01$ (two-tailed) were considered statistically significant to reduce the chance of type one error. For analyses of variance (ANOVAs) and multiple regression analyses, probabilities of $p < .05$ (two-tailed) were considered statistically significant in order to reduce the chance of Type II error and to be more conservative.

Results

Figure 1 shows participation and completion rates of all study groups, starting (on top) with the number of patients screened and included in the study, and ending (at the bottom) with the number of patients assessed per study group. Basic demographics and clinical characteristics for patients of Experiments 1 and 2 and for nonparticipants are provided in Table 2. For both studies there were few test scores identified as outliers: 2.6% (27) of all scores for Experiment 1, and 1.8% (40) for Experiment 2. Number of outliers per measure can be derived from the number of participants (n) as reported in Tables 3 and 4.

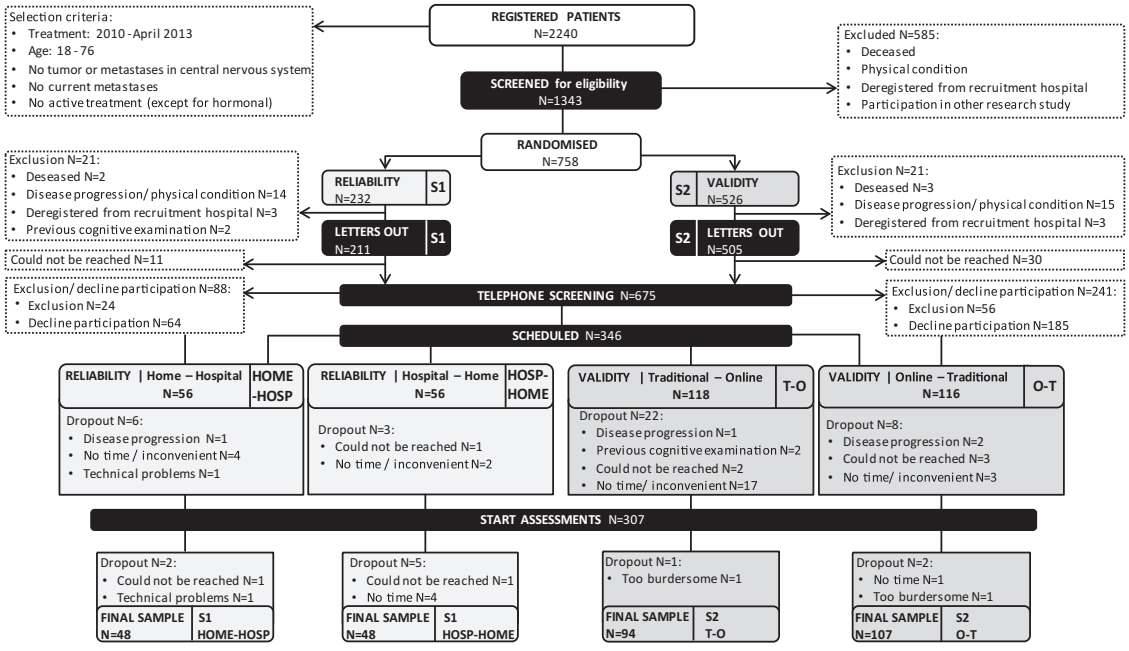

Figure 1. Participation and completion rates of all study groups.

Table 2. Demographics and clinical characteristics of patients in Experiments 1 and 2, and of nonparticipants.

	S1 (*N* = 96)	S2 (*N* = 201)	Nonparticipants (*N* = 419)
Gender			
Female, *n* (%)	57 (59)	112 (56)	274 (66)
Male, *n* (%)	39 (41)	89 (44)	144 (34)
Age (years), *M* (*SD*)	51.8 (11.9)	53.5 (12.3)	53.1 (13.5)
Education level[a]			N.A.[b]
Low, *n* (%)	1 (1)		
Medium, *n* (%)	41 (43)	78 (39)	
High, *n* (%)	54 (57)	123 (61)	
Comp. experience (years), *M* (*SD*)	18.4 (8.2)	18.5 (8.8)	N.A.[b]
IQ estimate, *M* (*SD*)	106 (10)	106.5 (9.6)	N.A.[b]
Tumor type			
Breast, *n* (%)	39 (41)	82 (41)	187 (44.5)
Testis/prostate, *n* (%)	20 (21)	39 (19)	69 (16.5)
Other, *n* (%)	37 (38)	84 (42)	162 (39)
Treatment type			
Surgery, *n* (%)	67 (70)	137 (68)	299 (71)
Chemotherapy, *n* (%)	77 (80)	154 (77)	323 (77)
Radiotherapy, *n* (%)	77 (80)	151 (75)	296 (71)
Hormonal therapy, *n* (%)	45 (47)	92 (46)	179 (43)
Immunotherapy, *n* (%)	8 (8)	28 (14)	57 (14)

Note. S1 = study group from Experiment 1 on test-retest reliability; S2 = study group from Experiment 2 on concurrent validity; *M* = mean; *SD* = standard deviation; Comp. experience = computer experience.
[a]Education is based on Verhage education scores 1–7 (Verhage, 1964), corresponding with the following U.S. years of education: 1: 1–5 years; 2: 6 years; 3: 7–8 years; 4: 7–9 years; 5: 7–10 years; 6: 7–16 years; 7: 17–20 years. Low = Verhage 1 or 2; Medium = Verhage 3, 4, or 5; and high = Verhage 6 or 7. [b]Not applicable; information is not available for this group.

Experiment 1: Effect of test setting and test–retest reliability

For Experiment 1, letters were sent to 211 patients, of whom 112 (53%) agreed to participate. Of these 112 patients, a total of 96 (86%) completed the study (59% female; M_{age} = 51.8 years, *SD* = 11.9). Nine patients dropped out before they started the study and six during the study (see Figure 1 for details). The data of one patient were excluded from analyses because of technical problems during several tests. In addition, one participant was excluded from analyses on the Word Learning test, as he indicated to have used a notepad during the home assessment.

The online assessment took about 1 hour and 15 minutes (on average 73 minutes for the first assessment) to complete. The test interval between the first and the second assessment was on average

Table 3. Test–retest reliability by means of ICC corrected for setting, Spearmans's ρ/Pearson's *r* (P), and literature on traditional tests.

Test	*n*	ICC setting[a]	Spearman's ρ/ Pearson's *r* (P)	Literature[b]
Connect the Dots I	90	.66***	.58***	.55 to .73
Connect the Dots II	92	.71***	.79***	.56 to .79
Wordlist Learning	95	.60***	.77***	.80
Wordlist Delayed recall	95	.49***	.72***	.83
Wordlist Recognition	95	.76***	.53***	.48
Reaction Speed	95	.52***	.35***	.20 to .82
Place the Beads	96	.40***	.48***	.38 to .70
Box Tapping	89	.29**	.29** (P)	.42 to .79
Fill the Grid	93	.33**	.36***	.72 to .86
Digit Sequences I	96	.66***	.66*** (P)	.61 to .78
Digit Sequences II	96	.64***	.64*** (P)	.46 to .71
Online total score	83	.78***	.78*** (P)	

[a]Intraclass correlation coefficient (ICC) controlled for setting. [b]Range of Pearson's r correlation coefficients as found in the literature on studies with retest intervals between 2 weeks and 6 months on adults without progressive disease is presented: Trail Making Test A (Bornstein, Baker, & Douglass, 1987; Bruggemans, Van de Vijver, & Huysmans, 1997; Silverstein et al., 2010; Spreen & Strauss, 1998); Trail Making Test B (Bornstein et al., 1987; Bruggemans et al., 1997; Strauss, Sherman, & Spreen, 2006); 15 Word Test learning & delayed recall (Saan & Deelman, 1986); 15 Word Test recognition (Andreotti & Hawkins, 2015); Visual Reaction Time FePsy (Andreotti & Hawkins, 2015; Lemay, Bedard, Rouleau, & Tremblay, 2004); Tower of London (Lemay et al., 2004; Lowe & Rabbitt, 1998; Schnirman, Welsh, & Retzlaff, 1998; Strauss et al., 2006); Corsi Block-Tapping (Lowe & Rabbitt, 1998; Miller, 1995; Nys et al., 2005; Silverstein et al., 2010); Grooved Pegboard (Dikmen et al., 1999; Ruff & Parker, 1993); Wechsler Adult Intelligence Scale (WAIS) Digit Span (Ryan, Lopez, & Paolo, 1996; Wechsler, 1997)
**$p < .01$. ***$p < .001$.

Table 4. Concurrent validity: Spearman or Pearson (P) correlation coefficients between the online tests and their traditional counterparts.

Test	*N*	Spearman's ρ/ Pearson's *r* (P)	Spearman's ρ/ Pearson's % of ceiling consistency[a]
Connect the Dots I	195	.57***	88 (.65)
Connect the Dots II	196	.70***	89 (.79)
Wordlist Learning	200	.64*** (P)	82 (.78)
Wordlist Delayed Recall	200	.59***	77 (.77)
Wordlist Recognition	200	.46***	92 (.50)
Reaction Speed	191	.49***	109 (.45)
Place the Beads[b]	189	.42***	72 (.58)
Box Tapping	194	.36***	75 (.48)
Fill the Grid	189	.45***	80 (.56)
Digit Sequences I	200	.43*** (P)	60 (.72)
Digit Sequences II	200	.52*** (P)	78 (.67)
Online total score[b]	171	.78*** (P)	

[a]√(Pearson's *r* traditional test based on literature × Pearson's *r*/Spearman's ρ online test based on Experiment 1).
[b]Analyses performed on data from the adapted version of Place the Beads.
***$p < .001$.

6.5 weeks (M = 45.5 days; range: 24 – 90). One participant had a test interval of 199 days; however, these data did not influence our results.

We did not find any differences on demographics and clinical characteristics (gender, age, computer experience, IQ, and tumor type) between the HOME-HOSP and HOSP-HOME subgroups.

Test setting

Two-way repeated measures ANOVAs indicated an interaction effect of "Subgroup × Time" for three of the 12 neuropsychological outcome measures: Reaction Speed ($p < .01$), Fill the Grid ($p < .01$), and the online total score ($p < .01$; Appendix B). On these tests, the HOME-HOSP and the HOSP-HOME subgroups showed reverse scoring patterns over time (carryover effect), which is illustrated in Figure 2(a) based on Fill the Grid scores. This effect was predominantly driven by test setting, as patients performed distinctly faster from the hospital than from the home setting on both time points (see Figure 2(b)). Figure 3 illustrates differences in mean scores for all neuropsychological tests (first assessment) between the home and the hospital setting, showing that patients performed generally better from the hospital setting.

Main effects for "subgroup" and "time" were examined for measures that did not show significant interaction effects. There were no significant main effects of subgroup on test performance (Appendix B). Positive main effects of time were found for Connect the Dots II ($p < .05$), Wordlist Learning ($p < .01$), Wordlist Delayed Recall ($p < .01$), Wordlist Recognition ($p < .05$), and Place the Beads ($p < .01$).

Test–retest reliability

Test–retest reliability is reported in Table 3. As test setting proved to have a significant effect on three of the 12 outcome measures, we controlled for setting by entering its variance estimates in the main ICC calculations (see "ICC setting" in Table 3). Questionnaire results (HADS and MFI–20) are presented separately in Appendix C. The ICCs of the individual tests were all statistically significant ($p < .01$) and ranged from .29 (Box Tapping) to .76 (Wordlist Recognition). In addition, the online total score had an ICC of .78. Five tests had ICCs of below our ≥.60 benchmark: Box Tapping (ICC = .29), Fill the Grid (.33), Place the Beads (.40), Wordlist Delayed Recall (.49), and Reaction Speed (.52). We generally found Spearman and Pearson correlations comparable to the highest test–retest correlations of the traditional tests reported in the literature from studies with similar test intervals, with the exception of Fill the Grid, which has distinct lower test–retest correlations than those in the literature (.36 vs. .72 to .86; Dikmen, Heaton, Grant, & Temkin, 1999; Ruff & Parker, 1993). Furthermore, our Spearman/Pearson results were generally comparable with our ICC results (six measures below the .60 criterion), except for higher reliability for Wordlist Delayed Recall and lower reliability for Wordlist Recognition (see Table 3).

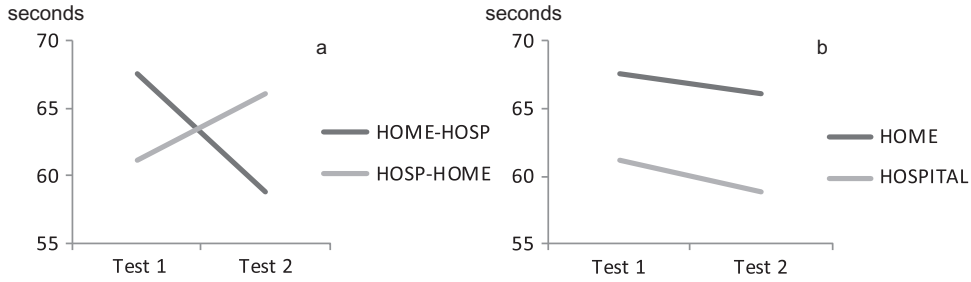

Figure 2. Influence of setting for performance on Fill the Grid test; illustration of a carryover effect. The *y*-axes indicate the time in seconds needed to complete the test; fewer seconds indicate better performance. (a) Mean performance on Fill the Grid over time per subgroup; the subgroup that first took the online test from home and then from the hospital (HOME-HOSP) improved over time whereas the subgroup that first took the online test from the hospital and then from home (HOSP-HOME) declined over time. (b) Mean performance on Fill the Grid over time, per setting. Performance from the hospital was faster than performance from home on both time points. Performance improved over time in both settings (home and hospital)..

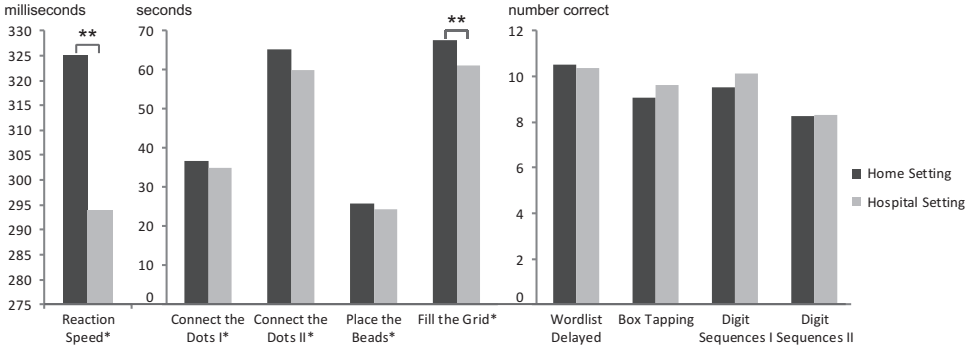

Figure 3. Home and hospital test scores on assessment 1. The *y*-axes represent (from left to right) milliseconds for Reaction Speed, seconds for Connect the Dots I & II, Place the Beads, and Fill the Grid, and number of correct responses for Wordlist Delayed Recall, Box Tapping, and Digit Sequences I & II. *Negative scoring scale (lower scores represent better performance). **Significant interaction effect "Subgroup × Time.".

Appendix D depicts both regular ICCs (not corrected for setting) and Pearson's *r* in order to indicate the proportion of systematic error (including practice effects), which is – in accordance with the ANOVA results – most pronounced for Wordlist Learning, Wordlist Delayed Recall, and Place the Beads.

Experiment 2: Concurrent validity

For Experiment 2, letters were sent to 505 patients, of whom 234 (46%) agreed to participate. Of these 234 patients, a total of 201 (86%) completed the study (56% female; M_{age} = 53.5 years; *SD* = 12.3). Thirty patients dropped out before they started the study and three during the study (see Figure 1). We did not find any differences between subgroups T-O (traditional tests first) and O-T (online tests first) on demographics and clinical characteristics.

Note that the Dutch reading test for adults could not be analyzed in five cases since four participants were non-native Dutch speakers (even though their Dutch language skills were sufficient to complete the assessments), and one participant had difficulties pronouncing words after mouth surgery. There were missing data on one of the online and one of the traditional assessments. One participant had to discontinue the online assessment after Wordlist Delayed Recall because of time constraints, resulting in missing data on Digit Sequences I & II and the questionnaires. From the traditional assessments, data were missing for one participant who did not complete the MFI–20 questionnaire. Furthermore, in comparison to Experiment 1, we used a slightly adapted version of the Place the Beads test for Experiment 2 in order to increase problem variability and decrease practice effects. Place the Beads was not analyzed for 10 participants who completed the old version. Data from the traditional verbal fluency test were not analyzed because the online equivalent of this test is still in development.

Concurrent validity

Table 4 shows concurrent validity results from Spearman or Pearson correlations between the online neuropsychological tests and their traditional

equivalents (questionnaire results are presented in Appendix C). Spearman/Pearson correlation coefficients were all significant ($p < .001$) and ranged from .36 to .70. The highest correlation was found for the online total score (Pearson's $r = .79$). Box Tapping was the only test that correlated with its traditional test below .40. In order to further interpret concurrent validity levels of the online tests, Table 4 also shows the correlations found in proportion to the ceiling consistency. Considering this maximum correlation that can be expected, concurrent validity was found to be medium to large. Digit Sequences I did have relatively low correlations with its traditional test (60% of the ceiling consistency). Although the validity of Box Tapping was the lowest of all tests, the correlation with its traditional test was 81% of the ceiling consistency. Figure 4 further illustrates the differences between the ceiling consistency and the concurrent validity found for the online tests.

Factors influencing online neuropsychological test scores

Possible influencing factors (i.e., "traditional test score," "age," "gender," "IQ," "computer skills," and "computer experience") of online neuropsychological test scores were included in the multivariate models. Computer experience (self-reported, in years) was found to be associated with only one of the outcome measures (Wordlist Delayed Recall) and was therefore excluded from the models.

Results from the multivariate analyses are presented in Table 5. Wordlist Recognition was not analyzed because there was hardly any variance in the data (ceiling effect). We found higher traditional test scores to be significantly associated ($p < .05$) with higher online scores for all but one of the tests (Fill the Grid). Younger participants showed higher online test scores than older participants on Connect the Dots I & II, Place the Beads, Fill the Grid, and Digit Sequences

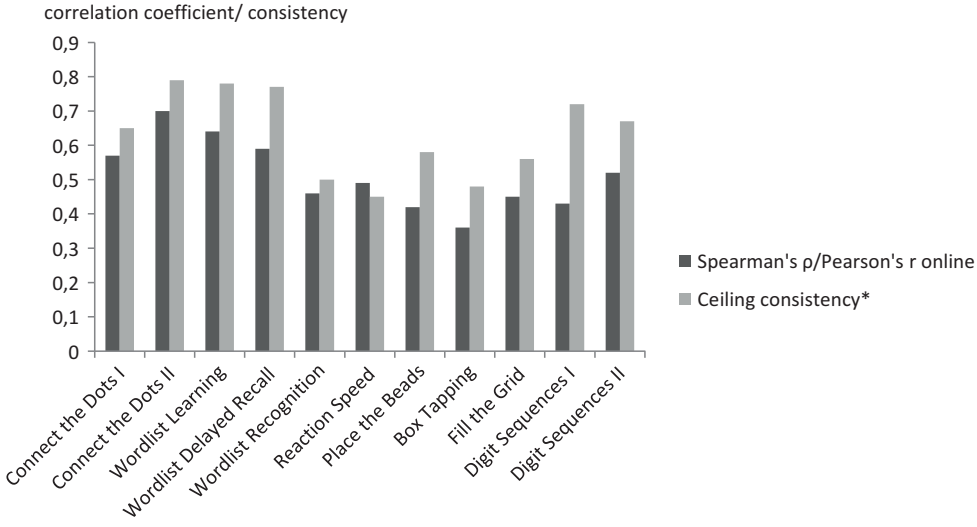

Figure 4. Correlations between traditional and online tests, compared to ceiling consistencies. *√(Pearson's *r* traditional test based on literature × Spearman's rho/Pearson's *r* online test based on Experiment 1).

Table 5. Multiple regression analyses; factors associated with online test scores.

	Traditional test		Age		Gender[a]		IQ		Computer skills		Model
Test	St. *B*	*SE B*	St. *B*	*SE B*	St. *B*	*SE B*	St. *B*	*SE B*	St. *B*	*SE B*	R^2
Connect the Dots I	0.27***	0.053	−0.29***	0.049	0.03	0.902	0.09	0.047	0.35***	1.301	.58
Connect the Dots II	0.42***	0.053	−0.35***	0.111	0.13**	1.980	0.09	0.111	0.16*	2.762	.61
Wordlist Learning	0.54***	0.063	−0.07	0.061	−0.13*	1.157	0.23***	0.060	0.01	1.554	.50
Wordlist Delayed	0.48***	0.073	−0.10	0.025	0.02	0.478	0.19**	0.025	0.04	0.647	.35
Reaction Speed	0.34***	0.102	−0.06	0.437	0.08	8.084	0.04	0.428	0.24**	11.419	.23
Place the Beads	0.31***	0.072	−0.20*	0.141	0.11	2.678	0.04	0.143	0.11	3.769	.22
Box Tapping	0.29***	0.087	−0.18	0.016	0.15*	0.296	0.05	0.016	0.20*	0.414	.26
Fill the Grid	0.07	0.057	−0.34***	0.070	0.15**	1.264	0.02	0.067	0.42***	1.710	.49
Digit Sequences I	0.39***	0.084	−0.02	0.016	0.11	0.316	0.00	0.018	0.19	0.434	.22
Digit Sequences II	0.43***	0.093	−0.26**	0.017	0.09	0.336	0.11	0.018	0.00	0.463	.31
Online total score	0.63***	0.052	−0.11	0.003	0.09	0.046	0.06	0.003	0.23***	0.063	.68

Note. *SE B* = standard error of beta; St. *B* = standardized beta; coefficients and *SE* values are only provided for the predictors that proved significant ($p < .1$) in the univariate analysis.
[a]Female is reference category.
*$p < .05$. **$p < .01$. ***$p < .001$.

II. Better computer skills were associated with higher online scores for Connect the Dots I & II, Reaction Speed, Box Tapping, Fill the Grid, and the online total score. Furthermore, female participants scored better than male participants on the online Wordlist Learning test, while male participants outperformed female participants on Connect the Dots II, Box Tapping, and Fill the Grid. Estimated IQ was least predictive for online test scores; participants with a higher IQ showed higher online test scores on Wordlist Learning & Delayed Recall. In seven out of 11 tests less than 50% of variance was explained by traditional test scores, age, gender, IQ, and computer skills (see R^2 levels, Table 5).

Since we found significant correlations between computer skills and age (ranging from r = .44 to .60), we did additional pathway (mediation) analyses to look into the effect of computer skills on online test scores, after being mediated by age (Hayes, 2012). Using bootstrapping procedures, unstandardized indirect effects were computed for each of 5000 bootstrapped samples with a 95% confidence interval. When mediating the effect of computer skills on all online test scores by age, the positive direct effect (path c') of computer skills was significant for all tests (see Table 6).

Discussion

The current studies assessed usability, test–retest reliability, and concurrent validity of the newly developed Amsterdam Cognition Scan, as well as the effect of test setting on performance, in order to indicate its usefulness to obtain reliable self-administered measures of a broad range of cognitive functions.

First, results of the usability of the ACS were positive. All participants were able to complete the tests from home and without the help of a test leader (see Appendix E for an overview of usability data).

With respect to test–retest reliability, Experiment 1 showed that the performance of six out of our 11 tests was sufficient, with ICCs exceeding the preestablished benchmark of .60. However, the five remaining tests (Wordlist Delayed Recall, Reaction Speed, Place the Beads, Box Tapping, and Fill the Grid) did not make this benchmark. To some extent, these suboptimal reliability scores may simply reflect the relatively low inherent test–retest reliability of the original tests that we based our online versions upon. Especially for the offline equivalents of Reaction Speed, Place the Beads, and Box Tapping, extant literature shows that test–retest reliabilities commonly do not exceed .60 (Pearson/Spearman).

Additionally, the fact that we conducted our study in patients rather than in healthy subjects may have reduced the test–retest reliability scores we observed. In patients, higher variance in health and psychological factors may be expected than in healthy participants, resulting in lower test–retest congruity. Indeed, collecting reference data from healthy subjects seems to improve test–retest reliability (Feenstra et al., in press). Furthermore, it should be noted that compared to tests of other online assessment tools, evaluated in healthy participants [Memoro (Hansen et al., 2016), with ICCs ranging from .55 to .74; BAM-COG (Aalbers, Baars, Olde Rikkert, & Kessels, 2013), with ICCs ranging from .17 to .65; NutriCog (Assmann et al., 2016), with Spearman correlations ranging from .36 to .65; and Brain on Track (Ruano et al., 2016), with ICCs between .51 and .82] the ACS does not underperform.

Arguably, one of the ACS's online tests, Fill the Grid, had particularly low reliability results compared to its offline counterpart. A possible explanation for this difference is the fact that we made rather extensive design changes in developing this particular online test. As a result, Fill the Grid may be relatively sensitive to

Table 6. The relationship between computer skills and online test scores mediated by age.

	Path c'		Path ab		Path ab‡		
	B	*SE*	*B*	*SE*	*B*	*SE*	CI
Connect the Dots I	−8.61***	1.28	−4.46***	0.89	−4.46	0.98	[−6.48, −2.65]
Connect the Dots II	−13.84***	2.96	−11.11***	2.08	−11.11	2.11	[−15.75, −7.36]
Wordlist Learning	4.09*	1.91	2.89*	1.17	2.89	1.12	[0.87, 5.37]
Wordlist Delayed	0.95*	0.37	0.54**	0.18	0.54	0.21	[0.22, 1.10]
Reaction Speed	−38.74*	11.82	−2.63	7.28	−2.63	7.54	[−16.80, 13.00]
Place the Beads	−10.36**	3.59	−3.28	2.16	−3.28	2.39	[−8.11, 1.36]
Box Tapping	1.12**	0.41	0. 558*	0.25	0.55	0.25	[0.11, 1.08]
Fill the Grid	−10.59***	1.62	−4.49***	1.07	−4.49	1.15	[−6.91, −2.38]
Digit Sequences I	1.30**	0.44	−0.16	0.27	−0.16	0.27	[−0.72, 0.35]
Digit Sequences II	0.98*	0.50	0.54	0.30	0.54	0.29	[−0.02, 1.14]
Online total	0.38***	0.08	0.21**	0.05	0.21	0.05	[0.11, 0.32]

Note. B = unstandardized beta; CI = 95% confidence interval. Path c' = direct effect of computer skills on test score. Path ab = Indirect effects of computer skills on test score through age (normal theory). Path ab‡ = indirect effects of computer skills on test score through age (bootstrap).
*p < .05. **p < .01. ***p < .001.

variance related to the online mode, such as computer configuration. Alternatively, between-subject variability in practice effects could be a possible explanation for this weak test–retest reliability. For participants lower in computer skills, the relative complex drag-and-drop technique (Hansen, Haferstrom, Brunner, Lehn, & Haberg, 2015) needed to complete this test is likely to improve over assessments (Lezak et al., 2004), while for more skilled participants test performance is likely to be more constant. Even though there are plausible explanations for the relatively low test–retest reliability on Fill the Grid, our results suggest that this test could be improved.

Our test–retest reliability analyses furthermore showed that the online test battery as a whole does produce quite consistent results. Although perhaps unsurprising, as averaged scores generally yield higher correlations than single measurements (Andreotti & Hawkins, 2015; de Vet et al., 2011), the highest test–retest reliability (an ICC of .78) was found for the online total score.

The currently observed test–retest reliabilities apply to testing with similar intervals (several weeks). Such intervals are realistic for clinical studies (Lin et al., 2013), but it would be valuable to know whether our observed reliability levels are similar for shorter or longer retest intervals. Therefore, we propose the use of different intervals in future studies on the ACS. Another issue that future studies should address is that of practice effects. In the current study, direct analysis of practice effects was hampered by our counterbalanced design and the possibly instable patient population. To be able to study practice effects better, we have recently completed collecting data from 249 healthy Dutch adults who were tested twice in the same (home) setting.

In Experiment 2, we found medium (>.30) correlations between online scores from the ACS and scores from their offline counterparts for six tests, and large (>.50) correlations for five tests (Cohen, 1988). The only correlation below .40 was for Box Tapping (ρ = .36). These results are similar to findings from three other concurrent validity studies comparing self-administered online cognitive test batteries with traditional tests. Aalbers and colleagues (2013) found correlations ranging from .20 to .67 (Spearman) for the BAM-COG, Hansen and colleagues (2015) found correlations of .49 to .63 (Pearson) for Memoro, and Morrison and colleagues (2015) found correlations between .47 and .71 (Pearson) for the NeuroCognitive Performance Test.

To understand why not all correlations between our online tests and their offline counterparts were large, it is important to realize that perfect correlations are generally not expected when a new test has noteworthy differences compared to the original. Obviously, all our online tests are different in being computerized and conducted via the internet instead of face to face, and for some tests additional adaptations were done in order to enhance test usability and validity. As could be expected, we found lower concurrent validity when we introduced more adaptations. In addition, concurrent validity is bound by the tests' test–retest reliabilities (ceiling consistency), which in some cases are not very high (see Table 3). After taking the inherent test–retest reliabilities of the new online and the traditional tests into account, only Digit Sequences I has a relatively low validity score. This may be explained by the combination of a relatively high ceiling consistency and significant adaptations for online use; adaptations included verbal instead of visual stimulus presentation and typed instead of verbal responses.

Box Tapping was the only test that correlated below .40 (Spearman) with its traditional counterpart, but when taking ceiling consistency into account, its consistency with traditional test scores was relatively high (81%).

Highest concurrent validity was found for the online total score (ρ = .78), indicating that the battery as a whole produces valid measurement of a variety of cognitive functions.

To identify tests that need improvement, low reliability results will provide the main indication, as validity results may be affected by low reliability results and since relatively low validity results may be expected when large adaptations are made. This means that tests with below-benchmark reliability results based on this patient population (Box Tapping, Place the Beads, Fill the Grid, Wordlist Delayed Recall, and Reaction Speed) will need extra attention when optimizing the ACS.

Even though our study and previous studies indicate that self-administered online testing is feasible and functional to elicit valid and reliable cognitive measurements (Assmann et al., 2016; Cromer et al., 2015), it remains important to look into limitations associated with online testing, and assess and, when possible, reduce the influence of contextual factors. Our results did indicate that test setting can influence test performance. Generally, participants performed somewhat better when online assessments took place at the hospital than when they took place at home. However, only for two tests (Reaction Speed and Fill the Grid) and for the online total score was performance significantly better when assessed in the hospital. So far, only few studies have been published on the influence of test setting on cognitive measurements, and these showed inconsistent results (Kuhn & Solomon, 2014; Raz, Bar-

Haim, Sadeh, & Dan, 2012). We argue that better performance in the hospital setting could be due to the corresponding degree of structure. Despite clear instructions on optimal test environment and preparation, participants are more likely to get distracted at home than at the hospital. A complicating factor in the interpretation of the effects of setting on test performance is the variety of computer configurations present in the home setting. In the current studies, possible effects of computer configurations were difficult to quantify, as there was little variance in use of browsers, operating systems, and input devices at home, and hardly any variance (for the sake of standardization) in the hospital. Because of the larger variance at home, home assessments may be expected to result in somewhat lower average test scores than hospital or lab-based assessments, especially for speed-based tests, which depend on very precise time measurements (Kuhn & Solomon, 2014; Raz et al., 2012). Therefore, for comparative cognitive studies, we advise using one constant test setting: either home or lab based. This also means that, in order to provide representative norms for future studies, test setting must be taken into account when collecting reference data (Bauer et al., 2012; Hansen et al., 2015).

Explorative analyses on factors influencing the online test results indicate, apart from an association with age, IQ, and—to a lesser extent—gender, that performance on several ACS tests can be influenced by level of computer skills. Multiple regression analyses showed that computer skills had a positive effect on online test scores for all speed-based measures. Interestingly, age is a confounder, although our computer skill measures remain to be associated with cognitive performance after correction for age. Even though computer skills were not associated with any of the traditional cognitive measures, it is not clear to what extent our computer skills measure can be disentangled from cognitive functioning. Therefore, further analyses on the influence of demographic variables (age and gender), computer configuration, and traditional cognitive test scores should indicate whether and how results from computer skill tests can reliably be used to better interpret cognitive results.

Limitations

There are two limitations of the current studies that have not been addressed yet. First, our sample represents highly educated patients, a common phenomenon in cognitive research. This may limit generalizability of our results to lower educated groups. As ceiling effects—generally resulting in high test–retest coefficients—are more likely to take place in highly educated participants, we could have overestimated reliability of some of the tests of the ACS. As a second limitation, it is possible that patients with relative good computer skills were more likely to participate in this study. Patients with poor computer skills or computer anxiety may have been prone to decline participation in an online study. This may affect the generalizability of the study results to less skilled participants.

Conclusions

The Amsterdam Cognition Scan is one of the first self-administered online neuropsychological test batteries with thoroughly studied psychometric properties. The studies described in this paper provide first evidence for the utility of this new online test battery for efficient large-scale cognitive data collection. Results of application in a sample of adult cancer patients indicate high usability and encouraging reliability for measuring functioning in various cognitive ability areas. However, several tests will need to be optimized in order to improve reliability and to warrant reliable measures of all cognitive ability areas covered by the ACS.

At the moment, development of the ACS as an online portal is an ongoing process. Currently tests are being optimized, the number of tests is being expanded, parallel versions are being developed, an English version of the entire battery is being validated, and reference data have been, and will continue to be, collected. After optimization and further development, the ACS will be applicable broadly throughout the field of oncology, but also beyond.

Acknowledgments

First, we thank all participants who were willing to take the time and effort to complete the assessments. Moreover, thanks to Nick Daems for programming the test portal and Vincent van der Noort for developing the software tool for randomization. Finally, many thanks to Anna Meijer, Daphne Kunst-Stevens, Fleur van Ierschot, Frederique Bambach, and Laura Menschaart for their great help with data gathering and data entry.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Dutch Cancer Society, KWF Kankerbestrijding [grant number NKI 2010-4876].

References

Aalbers, T., Baars, M. A., Olde Rikkert, M. G., & Kessels, R. P. (2013). Puzzling with online games (BAM-COG): Reliability, validity, and feasibility of an online self-monitor for cognitive performance in aging adults. *Journal of Medical Internet Research*, *15*(12), e270. doi:10.2196/jmir.2860

Alpherts, W., & Aldenkamp, A. P. (1995). *FePsy: The iron psyche, manual*. Heemstede: Instituut voor Epilepsiebestrijding.

Anastasi, A. (1998). *Psychological testing* (6th ed.). New York: Macmillan.

Andreotti, C., & Hawkins, K. A. (2015). RBANS norms based on the relationship of age, gender, education, and WRAT-3 reading to performance within an older african american sample. *The Clinical Neuropsychologist*, *29*(4), 442–465. doi:10.1080/13854046.2015.1039589

Assmann, K. E., Bailet, M., Lecoffre, A. C., Galan, P., Hercberg, S., Amieva, H., & Kesse-Guyot, E. (2016). Comparison between a self-administered and supervised version of a web-based cognitive test battery: Results from the NutriNet-Sante cohort study. *Journal of Medical Internet Research*, *18*(4). doi:10.2196/jmir.4862

Barak, A., & Buchanan, T. (2003). Internet-based psychological assessment. In R. Kraus, J. S. Zack, & G. Stricker (Eds.), *Online counseling: A handbook for mental health professionals* (pp. 217–240). San Diego, CA: Elsevier Academic Press.

Barak, A., & English, N. (2002). Prospects and limitations of psychological testing on the internet. *Journal of Technology in Human Services*, *19*, 65–89. doi:10.1300/J017v19n02_06

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. doi:10.1080/13854046.2012.663001

Bilder, R. M. (2011). Neuropsychology 3.0: Evidence-based science and practice. *Journal of the International Neuropsychological Society*, *17*(1), 7–13. doi:10.1017/S1355617710001396

Birnbaum, M. H. (2004). Human research and data collection via the internet. *Annual Review of Psychology*, *55*, 803–832. doi:10.1146/annurev.psych.55.090902.141601

Bornstein, R. A., Baker, G., & Douglass, A. (1987). Short-term retest reliability of the halstead-reitan battery in a normal sample. *The Journal of Nervous and Mental Disease*, *175*(4), 229–232. doi:10.1097/00005053-198704000-00007

Bruggemans, E. F., Van de Vijver, F. J., & Huysmans, H. A. (1997). Assessment of cognitive deterioration in individual patients following cardiac surgery: Correcting for measurement error and practice effects. *Journal of Clinical and Experimental Neuropsychology*, *19*(4), 543–559. doi:10.1080/01688639708403743

Brünken, R., & Plass, J. L. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, *38*(1), 53–61. doi:10.1207/S15326985EP3801_7

Buchanan, T., & Smith, J. L. (1999). Using the internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology*, *90*(1), 125–144. doi:10.1348/000712699161189

Caine, C., Mehta, M. P., Laack, N. N., & Gondi, V. (2012). Cognitive function testing in adult brain tumor trials: Lessons from a comprehensive review. *Expert Review of Anticancer Therapy*, *12*(5), 655–667. doi:10.1586/era.12.34

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Cole, W. R., Arrieux, 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*, *28*(7), 732–742. doi:10.1093/arclin/act040

Cromer, J. A., Harel, B. T., Yu, K., Valadka, J. S., Brunwin, J. W., Crawford, C. D., … Maruff, P. (2015). Comparison of cognitive performance on the Cogstate brief battery when taken in-clinic, in-group, and unsupervised. *The Clinical Neuropsychologist*, *29*(4), 542–558. doi:10.1080/13854046.2015.1054437

Culbertson, W. C., & Zillmer, E. A. (2001). *Tower of london-drexel (TOL-DX) technical manual*. North Tonawanda, NY: Multi-Health Systems.

Darby, D. G., Fredrickson, J., Pietrzak, R. H., Maruff, P., Woodward, M., & Brodtmann, A. (2014). Reliability and usability of an internet-based computerized cognitive testing battery in community-dwelling older people. *Computers in Human Behavior*, *30*, 199–205. doi:10.1016/j.chb.2013.08.009

de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine: A practical guide* (2nd ed.). Cambridge: Cambridge University Press.

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika*, *62*(3), 531–545. doi:10.1093/biomet/62.3.531

Dikmen, S. S., Heaton, R. K., Grant, I., & Temkin, N. R. (1999). Test-retest reliability and practice effects of expanded Halstead-Reitan neuropsychological test battery. *Journal of the International Neuropsychological Society*, *5*(4), 346–356. doi:10.1017/S1355617799544056

Dougherty, J. H., Jr., Cannon, R. L., Nicholas, C. R., Hall, L., Hare, F., Carr, E., … Arunthamakun, J. (2010). The computerized self test (CST): An interactive, internet accessible cognitive screening test for dementia. *Journal of Alzheimer's Disease*, *20*(1), 185–195. doi:10.3233/JAD-2010-1354

Feenstra, H. E. M., Vermeulen, I. E., Murre, J. M. J., & Schagen, S. B. (in press). *Reference data and reliability for the Amsterdam cognition scan: Reports on a new online tool for self-administered cognitive testing*.

Gates, N. J., & Kochan, N. A. (2015). Computerized and online neuropsychological testing for late-life cognition and neurocognitive disorders: Are we there yet? *Current Opinion in Psychiatry*, *28*(2), 165–172. doi:10.1097/YCO.0000000000000141

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19*(5), 847–857. doi:10.3758/s13423-012-0296-9

Gunter, B., Nicholas, D., Huntington, P., & Williams, P. (2002). Online versus offline research: Implications for evaluating digital media. *Aslib Proceedings*, *54*(4), 229–239. doi:10.1108/00012530210443339

Hansen, T. I., Haferstrom, E. C., Brunner, J. F., Lehn, H., & Haberg, A. K. (2015). Initial validation of a web-based self-administered neuropsychological test battery for older adults and seniors. *Journal of Clinical and Experimental Neuropsychology*, *37*(6), 581–594. doi:10.1080/13803395.2015.1038220

Hansen, T. I., Lehn, H., Evensmoen, H. R., & Håberg, A. K. (2016). Initial assessment of reliability of a self-administered web-based neuropsychological test battery. *Computers in Human Behavior*, *63*, 91–97. doi:10.1016/j.chb.2016.05.025

Hayes, A. F. (2012). PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling [White paper]. Retrieved from http://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/SobelTest

Internet world stats. (2017). *World internet usage and population statistics*, Retrieved from http://www.internetworldstats.com/stats.htm

ImPACT Applications. Inc.. (n.d.). *Frequently asked questions - Can ImPACT be taken at home?* Retrieved May, 2016, from https://www.impacttest.com/about/?Frequently-Asked-Questions-7

Iverson, G. L., Brooks, B. L., Ashton, V. L., Johnson, L. G., & Gualtieri, C. T. (2009). Does familiarity with computers affect computerized neuropsychological test performance? *Journal of Clinical and Experimental Neuropsychology*, *31*(5), 594–604. doi:10.1080/13803390802372125

Kessels, R. P. C., van Zandvoort, M. J. E., Postma, A., Kappelle, L. J., & Edward, H. (2010). The corsi block-tapping task: Standardization and normative data. *Applied Neuropsychology*, *7*(4), 37–41.

Kløve, H. (1963). *Grooved pegboard test user instructions*. Lafayette: Lafayette Instruments.

Kuhn, A. W., & Solomon, G. S. (2014). Supervision and computerized neurocognitive baseline test performance in high school athletes: An initial investigation. *Journal of Athletic Training*, *49*(6), 800–805. doi:10.4085/1062-6050-49.3.66

Lemay, S., Bedard, M., Rouleau, I., & Tremblay, P. G. (2004). Practice effect and test-retest reliability of attentional and executive tests in middle-aged to elderly subjects. *The Clinical Neuropsychologist*, *18*(2), 284–302.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766. doi:10.1016/j.jesp.2013.03.013

Lezak, M. D., Howieson, D. B., Loring, D. W., Hannay, H. J., & Fischer, J. S. (2004). *Neuropsychological assessment* (4th ed.). Oxford: Oxford University Press.

Lin, N. U., Lee, E. Q., Aoyama, H., Barani, I. J., Baumert, B. G., Brown, P. D., … Wen, P. Y. (2013). Challenges relating to solid tumour brain metastases in clinical trials, part 1: Patient population, response, and progression. A report from the RANO group. *The Lancet Oncology*, *14*(10), e396–e406. doi:10.1016/S1470-2045(13)70311-5

Littleton, A. C., Register-Mihalik, J. K., & Guskiewicz, K. M. (2015). Test-retest reliability of a computerized concussion test: CNS vital signs. *Sports Health: A Multidisciplinary Approach*, *7*, 443–447. doi:10.1177/1941738115586997

Lowe, C., & Rabbitt, P. (1998). Test\re-test reliability of the CANTAB and ISPOCD neuropsychological batteries: Theoretical and practical issues. *Neuropsychologia*, *36*(9), 915–923. doi:10.1016/S0028-3932(98)00036-0

Medalia, A., Lim, R., & Erlanger, D. (2005). Psychometric properties of the web-based work-readiness cognitive screen used as a neuropsychological assessment tool for schizophrenia. *Computer Methods and Programs in Biomedicine*, *80*(2), 93–102. doi:10.1016/j.cmpb.2005.06.007

Miller, E. (1995). CaLCAP psychometric properties. Retrieved March 27, 2017, from http://www.calcaprt.com/metrics.htm

Morrison, G. E., Simone, C. M., Ng, N. F., & Hardy, J. L. (2015). Reliability and validity of the neurocognitive performancetest, a web-based neuropsychological assessment. *Frontiers in Psychology*, *6*(1652). doi:10.3389/fpsyg.2015.01652

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological testing on the internet: New problems, old issues. *The American Psychologist*, *59*(3), 150–162. doi:10.1037/0003-066X.59.3.150

Nys, G., Van Zandvoort, M., De Kort, P., Jansen, B., Van der Worp, H., Kappelle, L., & De Haan, E. (2005). Domain-specific cognitive recovery after first-ever stroke: A follow-up study of 111 cases. *Journal of the International Neuropsychological Society*, *11*(07), 795–806. doi:10.1017/S1355617705050952

Parsey, C. M., & Schmitter-Edgecombe, M. (2013). Applications of technology in neuropsychological assessment. *The Clinical Neuropsychologist*, *27*(8), 1328–1361. doi:10.1080/13854046.2013.834971

Raz, S., Bar-Haim, Y., Sadeh, A., & Dan, O. (2012). Reliability and validity of the online continuous performance test among young adults. *Assessment*, *21*(1), 108–118. doi:10.1177/1073191112443409

Reips, U.-D. (2002a). Context effects in web surveys. In B. Batinic, U. D. Reips, & M. Bosnjak (Eds.), *Online Social Sciences* (pp. 95-101). Göttingen: Hogrefe & Huber.

Reips, U.-D. (2002b). Internet-based psychological experimenting: Five dos and five don'ts. *Social Science Computer Review*, *20*(3), 241–249. doi:10.1177/08939302020003002

Reips, U.-D. (2002c). Standards for internet-based experimenting. *Experimental Psychology*, *49*(4), 243–256. doi:10.1026/1618-3169.49.4.243

Reitan, R. M. (1958). Validity of the trail making test as an indicator of organic brain damage. *Perceptual and Motor Skills*, *8*, 271–276. doi:10.2466/pms.1958.8.3.271

Ruano, L., Sousa, A., Severo, M., Alves, I., Colunas, M., Barreto, R., … Bento, V. (2016). Development of a self-administered web-based test for longitudinal cognitive assessment. *Scientific Reports*, *6*. doi:10.1038/srep19114

Ruff, R. M., & Parker, S. B. (1993). Gender-and age-specific changes in motor speed and eye-hand coordination in adults: Normative values for the finger tapping and

grooved pegboard tests. *Perceptual and Motor Skills*, *76* (3c), 1219–1230. doi:10.2466/pms.1993.76.3c.1219

Ryan, J. J., Lopez, S. J., & Paolo, A. M. (1996). Temporal stability of digit span forward, backward, and forward minus backward in persons aged 75-87 years. *Cognitive and Behavioral Neurology*, *9*(3), 206–208.

Saan, R. J., & Deelman, B. G. (1986). *The 15-words test A and B*. Groningen: University of Groningen, Department of Neuropsychology.

Schagen, S. B., Klein, M., Reijneveld, J. C., Brain, E., Deprez, S., Joly, F., … Wefel, J. S. (2014). Monitoring and optimising cognitive function in cancer patients: Present knowledge and future directions. *European Journal of Cancer Supplements*, *12*(1), 29–40. doi:10.1016/j.ejcsup.2014.03.003

Schatz, P. (2010). Long-term test-retest reliability of baseline cognitive assessments using ImPACT. *The American Journal of Sports Medicine*, *38*(1), 47–53. doi:10.1177/0363546509343805

Schatz, P., & Browndyke, J. (2002). Applications of computer-based neuropsychological assessment. *The Journal of Head Trauma Rehabilitation*, *17*(5), 395–410. doi:10.1097/00001199-200210000-00003

Schmand, B., Bakker, D., Saan, R., & Louman, J. (1991). The Dutch reading test for adults: A measure of premorbid intelligence level. *Tijdschrift Voor Gerontologie En Geriatrie*, *22*(1), 15–19.

Schmand, B., Groenink, S. C., & van den Dungen, M. (2008). Letter fluency: psychometric properties and Dutch normative data. *Tijdschrift voor gerontologie en geriatrie*, *39*(2), 64–76. doi:10.1007/BF03078128

Schnirman, G. M., Welsh, M. C., & Retzlaff, P. D. (1998). Development of the Tower of London-revised. *Assessment*, *5*(4), 355–360. doi:10.1177/107319119800500404

Silverstein, S. M., Berten, S., Olson, P., Paul, R., Willams, L. M., Cooper, N., & Gordon, E. (2007). Development and validation of a World-Wide-Web-based neurocognitive assessment battery: WebNeuro. *Behavior Research Methods*, *39*(4), 940–949. doi:10.3758/BF03192989

Silverstein, S. M., Jaeger, J., Donovan-Lepore, A.-M., Wilkniss, S. M., Savitz, A., Malinovsky, I., … Marcello, S. (2010). A comparative study of the MATRICS and IntegNeuro cognitive assessment batteries. *Journal of Clinical and Experimental Neuropsychology*, *32*(9), 937–952. doi:10.1080/13803391003596496

Smets, E. M., Garssen, B., Bonke, B., & De Haes, J. C. (1995). The Multidimensional Fatigue Inventory (MFI) psychometric qualities of an instrument to assess fatigue. *Journal of Psychosomatic Research*, *39*(3), 315–325. doi:10.1016/0022-3999(94)00125-O

Spreen, O., & Strauss, E. (1998). *A compendium of neuropsychological tests*. New York: Oxford University Press.

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

Trustram Eve, C., & de Jager, C. A. (2014). Piloting and validation of a novel self-administered online cognitive screening tool in normal older persons: The cognitive function test. *International Journal of Geriatric Psychiatry*, *29*(2), 198–206. doi:10.1002/gps.3993

Tsushima, M., Tsushima, W., Tsushima, V., Lim, N., Madrigal, E., Jackson, C., & Mendler, M. H. (2013). Use of ImPACT to diagnose minimal hepatic encephalopathy: An accurate, practical, user-friendly internet-based neuropsychological test battery. *Digestive Diseases and Sciences*, *58*(9), 2673–2681. doi:10.1007/s10620-013-2668-z

Van de Weijer-Bergsma, E., Kroesbergen, E. H., Prast, E. J., & Van Luit, J. E. (2015). Validity and reliability of an online visual-spatial working memory task for self-reliant administration in school-aged children. *Behavioral Research Methods*, *47*(3), 708–719. doi:10.3758/s13428-014-0469-8

Van den Burg, W., Saan, R. J., & Deelman, B. G. (1985). *15-woordentest. Provisional manual.* Groningen, The Netherlands: University Hospital, Department of Neuropsychology.

van Hooijdonk, C., & Krahmer, E. (2008). Information modalities for procedural instructions: The influence of text, pictures, and film clips on learning and executing RSI exercises. *IEEE Transactions on Professional Communication*, *51*, 50–62. doi:10.1109/TPC.2007.2000054

Verhage, F. (1964). *Intelligentie en leeftijd: Onderzoek bij Nederlanders van twaalf tot zevenenzeventig jaar* (Doctoral thesis) Van Gorcum, Assen.

Wechsler, D. (1997). *WAIS-III: Wechsler adult intelligence scale*. San Antonio, TX: Psychological Corporation.

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research*, *19*(1), 231–240. doi:10.1519/15184.1

Zigmond, A. S., & Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica*, *67* (6), 361–370. doi:10.1111/acp.1983.67.issue-6

Appendices

Appendix 1. Online neuropsychological tests and computer skill tests of the Amsterdam Cognition Scan

Neuropsychological tests

(1) Connect the Dots I & II

I: This test measures visuomotor tracking and planning. Circles, numbered from 1 to 25, are presented on the screen. Participants are instructed to click the numbered circles in increasing order, as quickly as possible and without making mistakes.

II: In combination with Part I, this test measures cognitive flexibility and divided attention. Again, 25 circles are presented on the screen. But this time 13 of them are numbered, and 12 contain a letter (A to L). Participants are instructed to alternate clicking digits (in increasing numerical order) and letters (in alphabetic order), as fast as possible and without making mistakes.

For both Part I and Part II, whenever the wrong circle is clicked, the circle colors red, and participants have to continue clicking in the right sequence. Performance is scored by calculating the total time (from stimulus presentation to completion) for Part I and Part II. Higher scores on Parts I and II separately indicate slower, and therefore worse, performance.

(2a) Wordlist Learning

This test measures verbal learning. Participants are asked to remember a list of 15 unrelated words. The words have been selected for matching length and frequency with the words from the traditional equivalent test. The words are presented one by one on the screen. After seeing the last word from the list, an entry field appears in which participants can type all the words they remember. The words have to be separated by at least one space, but the order in which the words are entered is not important. Spelling mistakes are programmed to be ignored up to two characters difference with the target word (e.g., the word "bellooon" would still be considered as referring to the target word "balloon" and counted as correct). When all the remembered words are entered, participants click on the "NEXT" button, after which the wordlist is presented again until a fifth presentation. All five presentations have a different fixed word order. Performance is scored as total number of correctly entered words throughout the five trials.

(2b) Wordlist Delayed Recall & Recognition

These tests measure retention of verbal information. The delayed recall of the Wordlist test consists of a free recall and a recognition task. During the free recall, participants are asked to type all the words they can remember from the previously presented word list once more. Performance is scored by the total number of correctly entered words. During the subsequent recognition task, participants are presented with 45 words on the screen, appearing one at the time. For each word they have to indicate whether or not it was on the list. If a word is recognized as a word from the list, the "yes"-button has to be clicked. If the word is not recognized, the "no"-button has to be clicked. Every word requires a response in order to continue. Beside the 15 target words, there are 30 distractor words, which are matched to the target words on semantics, word length, and word frequency. Performance is scored by summing all correct responses (both correctly recognized target words and correctly identified distractor words).

(3) Reaction Speed

Reaction speed is measured on 30 consecutive trials. Participants start by pressing the mouse button and keeping it pressed down. After a short time (2.5 to 4 s), a white block appears on the screen after which, as fast as possible, the mouse button has to be released. The white block always appears on the same place on the screen. Whenever participants let go of the mouse button too early (less than 130 ms after the appearance of the white block), they are instructed to try again. Performance is scored by averaging all correct trials in between 130 and 3950 ms. Higher scores indicate slower, and therefore worse, performance.

(4) Place the Beads

This test measures planning and response inhibition. Participants are presented with pictures of two constructions, both consisting of three sticks on which three distinctive balls (a grey, a black, and a white one) are placed. The upper construction is fixed and has to be rebuilt in the lower construction. The aim is to use as few moves as possible to make the lower construction resemble the upper construction. The balls of the lower construction can be moved by dragging them from their current to the desired location. There are two rules presented: (a) Only the upper beads can be moved; and (b) no more balls can be placed on a stick than the stick can fit (one, two, or three balls). Rule-incongruent

moves are programmed to be impossible to perform. There are 10 problems to solve. If a problem is not solved within the 2-minute time limit, the test continues with the next problem until all 10 problems have been presented. Performance is scored by summing all moves additional to the minimum number of moves for each problem. If the time limit is reached, the maximum score of 20 is given.

(5) Box Tapping

This test measures visuospatial short-term memory. Participants are presented with a number of blocks on the screen. The blocks blink one by one. Participants are asked to remember the order in which the blocks blinked. When indicated, they must click on the blocks in the same order. After finishing the response, participants click on the "done"-button; they are then presented with the next series. After two correct series of the same length, the number of blocks is increased by one. This continues until there are 12 blocks on the screen, or whenever there are two incorrect responses within the same series length. Performance is scored by summing the total number of correctly repeated series.

(6) Fill the Grid

This test measures fine motor skills. A grid is presented that consists of 25 squares. Participants are instructed to fill the grid with red blocks as quickly as possible. As soon as the red block appears underneath the grid, it has to be dragged to the square in which a plus sign is shown. The blocks have to be positioned exactly within the edges of the square. Once a square has been filled correctly, a new red block appears, which also has to be placed as fast as possible. The task is completed as soon as all squares are filled. There is a time limit of 3 minutes. Performance is scored by calculating the total time from first to last move. Higher scores indicate slower, and therefore worse, performance.

(7) Digit Sequences I & II

I: This test measures attention. A series of digits appear on the screen one by one. Participants have to remember the digits and the order in which they are presented. After the last digit, a digit bar appears on which the digits have to be entered in the correct order. The response is made by clicking on the digits. The selected digits then appear above the bar in the order selected. Mistakes can be corrected by using the "delete"-button. The response is finished by clicking the "done"-button, after which the next series appears. After two correct series of the same length, an extra digit is added. The test starts with a series of three digits and halts either after the last series of 12 digits, or whenever there are two incorrect responses within the same series length. Performance is scored by summing all correctly repeated series.

II: This test measures working memory. The presentation of Part II is similar to that of Part I, but this time the digits have to be entered in reverse order. Performance is again scored by summing the total number of correctly repeated series.

Online neuropsychological tests of the Amsterdam Cognition Scan are shown in Figure A1.

Computer skill tests

(A) Type Skills

This test measures speed and accuracy of keyboard typing. A target sentence is presented, which has to be retyped in the response box below. Performance is scored by calculating the total time from first to last key press. Higher scores indicate slower, and therefore worse, performance.

(B) Click Skills

This test measures speed of precise clicking via the input device of choice (e.g., mouse or touchpad). A number of circles are presented on the screen, together forming a spiral. From the outside to the center of the spiral, the circles go from large to small. Participants are instructed to click the circles form the outer part of the spiral to the center of the spiral, as quickly as possible and without skipping any of the circles. Performance is scored by calculating the total time from first to last click. There is a time limit of 3 minutes. Higher scores indicate slower, and therefore worse, performance.

(C) Drag Skills

This test measures speed of precise drag-and-drop actions. For every trial (eight in total) two elements of the same shape are presented on the screen. One of the elements is black; the other element is white and slightly larger than the black one. Participants are instructed to drag the black element exactly into the white element. After succeeding in doing so, the next trial is presented. Performance is scored by calculating the total time from first to last action. There is a time limit of 3 minutes. Higher scores indicate slower, and therefore worse, performance.

Computer skill tests of the Amsterdam Cognition Scan are shown in Figure A2.

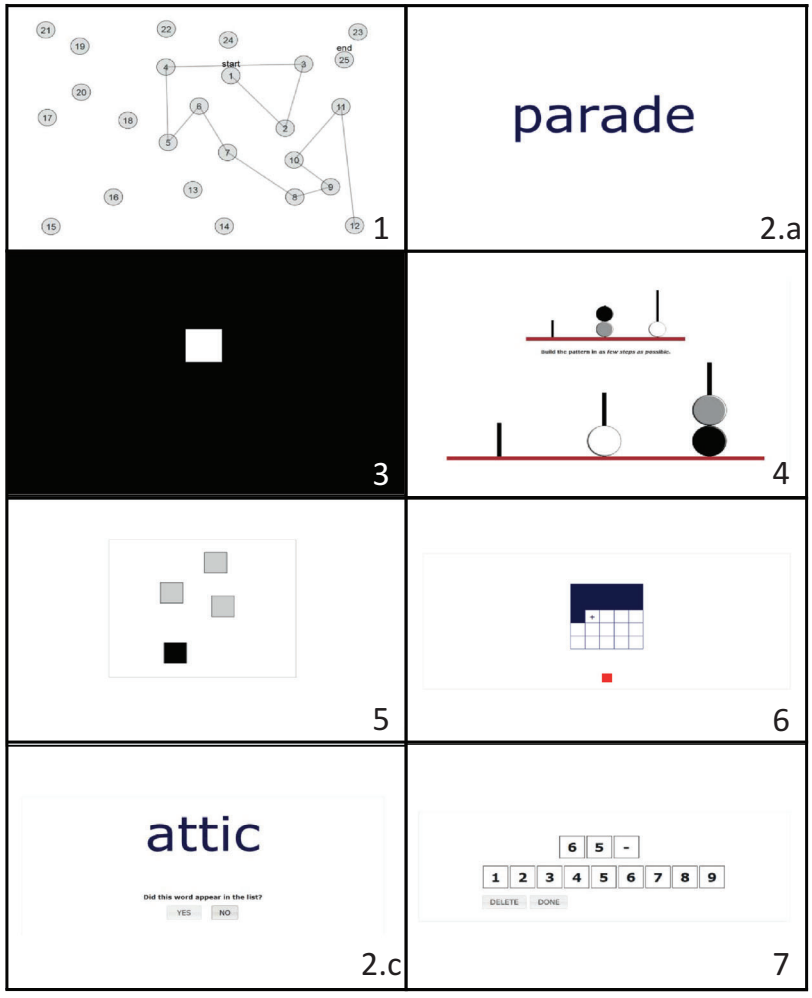

Figure A1. Online neuropsychological tests of the Amsterdam Cognition Scan.

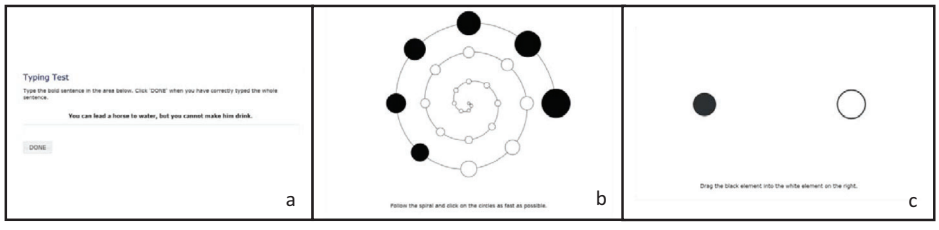

Figure A2. Computer skill tests of the Amsterdam Cognition Scan.

Appendix B. Test scores and results of two-way repeated measures ANOVAs on "Subgroup × Time," "Subgroup," and "Time."

Test	Time	Home setting *M* (*SD*)	Hospital setting	Subgroup × Time[a] *F*	Time *F*	Subgroup *F*
Connect the Dots I	1	36.6 (7.5)	34.9 (5.8)	3.12	3.38	0.07
	2	34.8 (7.9)	34 (7.1)			
Connect the Dots II	1	65.2 (19.9)	60 (17.6)	2.33	5.23*	0.38
	2	58.6 (18.1)	58.7 (18.1)			
Wordlist Learning	1	47.8 (10.8)	49.6 (8.5)	1.02	129.75**	0.35
	2	56.5 (7.9)	56.0 (8.1)			
Wordlist Delayed Recall	1	10.7 (2.5)	11.2 (2.5)	1.21	24.85**	0.07
	2	12.5 (2.2)	12.4 (2.3)			
Wordlist Recognition	1	43.9 (2.5)	43.9 (1.5)	0.62	5.22*	0.08
	2	44.1 (0.9)	44.3 (1.8)			
Reaction Speed	1	325 (41.7)	294 (37.8)	38.01**		
	2	317 (51.2)	290 (51.2)			
Place the Beads	1	25.7 (19.9)	24.4 (19.6)	0.71	15.73**	0.00
	2	19.2 (10.6)	17.7 (10.6)			
Box Tapping	1	9.0 (2)	9.6 (1.5)	3.97	0.00	0.30
	2	9.2 (1.9)	9.5 (1.2)			
Fill the Grid	1	67.6 (12.8)	61.2 (10.2)	16.23**		
	2	66.1 (11.9)	58.8 (7)			
Digit Sequences I	1	9.5 (2.3)	10.1 (2.1)	1.53	0.58	0.86
	2	10.0 (2)	9.9 (2.2)			
Digit Sequences II	1	8.3 (2.7)	8.3 (2.4)	0.55	1.70	0.03
	2	8.5 (2.8)	8.7 (2.7)			
Online total score	1	0.05 (0.44)	0.18 (0.43)	35.67**		
	2	0.003 (0.48)	0.2 (0.48)			

Note. ANOVA = analysis of variance.
[a]Interaction between "subgroup" and "time." This interaction reflects the influence of test setting.
*$p < .05$. **$p < .01$.

Appendix C. Questionnaire results on test–retest reliability and concurrent validity.

	n	ICC setting[a]	ICC	Pearson's *r*
Test–retest reliability				
Questionnaires				
HADS anxiety	96	.88***	.88***	.89***
HADS depression	96	.85***	.85***	.85***
MFI total	96	.86***	.86***	.85***
Concurrent validity				
Questionnaires				
HADS anxiety	200			.94***
HADS depression	200			.94***
MFI total	199			.96***

Note. HADS = Hospital Anxiety and Depression Scale; MFI = Multidimensional Fatigue Inventory; ICC = intraclass correlation coefficient.
[a]Controlled for setting.
***$p < .001$.

Appendix D. Test–retest reliability: ICCs and Pearson's *r* correlations to indicate systematic error (including practice effects).

Test	*n*	ICC	Pearson's *r*
Connect the Dots I	90	.61***	.62***
Connect the Dots II	92	.66***	.68***
Wordlist Learning	95	.60***	.78***
Wordlist Delayed recall	95	.49***	.59***
Wordlist Recognition	95	.76***	.78***
Reaction Speed	95	.44***	.45***
Place the Beads	96	.40***	.48***
Box Tapping	89	.29**	.29**
Fill the Grid	93	.31**	.32**
Digit Sequences I	96	.66***	.66***
Digit Sequences II	96	.64***	.64***
Online total score	83	.78***	.78***

Note. ICC = intraclass correlation coefficient.
**$p < .01$. ***$p < .001$.

Appendix E. Usability data from online debriefing on the (first) assessments with the Amsterdam Cognition Scan.

Question	T-O hospital (*n* = 93) % (frequency)	O-T hospital (*n* = 106) % (frequency)	HOME-HOSP home (*n* = 46) % (frequency)	HOSP-HOME hospital (*n* = 46) % (frequency)
Others present during assessment	1.1 (1)	0 (0)	6.5 (3)	2.2 (1)
Tasks clear after practice session	0 (0)	0.9 (1)	0 (0)	0 (0)
Received help from others	0 (0)	0 (0)	0 (0)	6.5 (3)
Used aid[a]	1.1 (1)	0.9 (1)	4.3 (2)	0 (0)
Were disrupted	1.1 (1)	6.6 (7)	13 (6)	0 (0)
By telephone	1.1 (1)	6.6 (7)	6.5 (3)	0 (0)
By house member/pet	0 (0)	0 (0)	6.5 (3)	0 (0)
Experienced technical problems	2.2 (2)	5.7 (6)	8.7 (4)	6.5 (3)
Internet connection	1.1 (1)	4.7 (5)	6.5 (3)	2.2 (1)
Hardware problem	1.1 (1)	0.9 (1)	2.2 (1)	4.3 (2)
Preference assessment type	(*n* = 87)	(*n* = 90)	(*n* = 39)	(*n* = 38)
Home online	75.9 (66)	60 (54)	82.6 (38)	81.6 (31)
Hospital paper and pencil	8 (7)	2.2 (2)	0 (0)	0 (0)
Hospital online	16.1 (14)	37.8 (34)	2.2 (1)	18.4 (7)

Note. T-O: 1. traditional, 2. online. O-T: 1. online, 2. traditional. HOME-HOSP: 1. home, 2. hospital. HOSP-HOME: 1. hospital, 2. home.
[a]In all cases this concerned paper and pencil use.