View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

Computers in Human Behavior 63 (2016) 91–97

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Initial assessment of reliability of a self-administered web-based neuropsychological test battery

T.I. Hansen [a, b, *], H. Lehn [c], H.R. Evensmoen [a, b], A.K. Håberg [a, b]

[a] *Department of Neuroscience, NTNU, Norwegian University of Science and Technology, Trondheim, Norway*
[b] *Department of Medical Imaging, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway*
[c] *Faculty of Medicine, NTNU, Norwegian University of Science and Technology, Trondheim, Norway*

ARTICLE INFO

Article history:
Received 21 December 2015
Received in revised form
4 May 2016
Accepted 5 May 2016
Available online 17 May 2016

Keywords:
Computerized
Intraclass correlation
Internet
Cognitive
Memory

ABSTRACT

Introduction: Web-based neuropsychological testing can be an important tool in meeting the increasing demands for neuropsychological assessment in the clinic and in large research studies. The primary aim of this study was to investigate practice effects and reliability of self-administered web-based neuropsychological tests in Memoro. Due to lack of consistent analysis and reporting of reliability in the literature, especially intraclass correlation coefficients (ICC), we highlight how using different ICC measures results in different estimates of reliability.
Method: 61 (31 females) participants (mean age 53.3 years) completed the Memoro tests twice with a median of 14 days between testing.
Results: Practice effects were detected for all cognitive measures ($d = 0.32–0.61$), most pronounced for memory measures. Reliability estimated using two-way random effects single measure absolute agreement ICC(2,1) were between 0.55 and 0.74. Two-way mixed effects average measure consistency ICC(3,2), ranged from 0.79 to 0.89. Reliability was highest for the processing speed task and lower for the memory tasks.
Conclusions: Memoro tests had test-retest reliability similar to that of traditional, computerized and web-based test batteries used clinically and in research. It is important to carefully choose and specify the ICC implemented, as ICC(2,1) and ICC(3,2) give different results and reflect reliability of different measures.

© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Today neuropsychological testing is moving from its traditional pen-and-paper format with an examiner to computers and even further to self-administrated web-based testing (Bilder, 2011; Resch, McCrea, & Cullum, 2013; Zygouris & Tsolaki, 2014). Web-based testing is gaining in popularity as it is flexible, inexpensive, and has no geographical boundaries (Darby, Fredrickson, Pietrzak, Maruff, & Woodward, 2014; Haworth & Harlaar, 2007). Web-based testing was embraced early by sports medicine and for concussion management due to its mobile and flexible nature (Erlanger et al., 2003).

Web-based tests are also being used in assessment of aging, mild cognitive impairment and dementia related cognitive changes (Darby et al., 2014; Dougherty et al., 2010; Trustram Eve & de Jager, 2014), other disease-related cognitive changes (Medalia, Lim, & Erlanger, 2005), and more generally in individual and potential large-scale assessments of cognitive functions (Silverstein et al., 2007). Web-based tests identify the same cognitive constructs as traditional tests, and are even preferred above traditional tests by persons who are tested (Hansen, Haferstrom, Brunner, Lehn, & Håberg, 2015). Web-based tools might facilitate large cohort studies and help meet the increasing demand for cognitive assessment associated with the aging population and accompanying cognitive changes and disease. The International Association of Gerontology and Geriatrics (IAGG) recently recommended that all individuals ≥ 70 years should have their memory tested at least once annually (Morley et al., 2015). Such a change in clinical practice will require large resources, and computerized or web-based testing is one way to meet the new demands.

It is imperative that web-based tests can document good

[*] Corresponding author. Department of Neuroscience, Faculty of Medicine, Pb. 8905, Norwegian University of Science and Technology, 7491 Trondheim, Norway.
E-mail address: torivar.hansen@ntnu.no (T.I. Hansen).

http://dx.doi.org/10.1016/j.chb.2016.05.025

0747-5632/© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

psychometric properties, as these are critical for the tests' usability both in the clinic and in research. While the reliability of computerized tests has been well documented, the test-retest reliability of web-based tests remains underexplored (Gates & Kochan, 2015; Zygouris & Tsolaki, 2014). Furthermore, reliability estimates are in some cases only available for the computerized version and not the web-based version. This is problematic as it can be argued that when the manner of administration changes, the test becomes a new test (Bauer et al., 2012).

A challenge in the literature on web-based testing is the lack of consistent assessment of test-retest reliability. Pearson's *r* or one of the several intraclass correlation coefficients (ICCs) are commonly reported, alternatively Spearman's *rho* if the statistical assumptions of the Pearson's *r* and ICC are not met. Since Pearson's *r* is insensitive to systematic error (e.g. practice effects or fatigue effects) using an ICC method has been recommended because it takes into account both the within-subject change and the systematic group change over time (Vaz, Falkmer, Passmore, Parsons, & Andreou, 2013). This recommendation is, however, not universally accepted as it can be argued that practice effects are not a flaw to be corrected, but rather a natural phenomenon (Rousson, Gasser, & Seifert, 2002). Adding further complexity to this picture, there exist multiple ICC variants with regard to both the computational processes and nomenclature (McGraw & Wong, 1996; Shrout & Fleiss, 1979), leading to a multitude of measures. These variations have caused confusion in the field. Furthermore, lack of specification of ICC variant used, and application of incorrect ICC variants in the literature can lead to misconstrued impressions of the reliability of the tests (Bruce, Echemendia, Meeuwisse, Comper, & Sisco, 2014; Krebs, 1986; Weir, 2005).

We have developed Memoro, a self-administered web-based neuropsychological test platform for assessment of memory and related cognitive functions. We have recently investigated concurrent validity (Hansen et al., 2015). However, no test is valid unless it is reliable.

The primary aim of this study was to assess practice effects and reliability, i.e. consistency of test results across administrations, of the self-administered web-based tests in Memoro. The second aim was to highlight how different variants of intraclass correlation results in different estimates of reliability. As noted above, the literature on reliability has not always been precise or correct in its use of ICC. This has significant impact on the reliability measurement as illustrated using test-retest data from Memoro.

2. Material and methods

2.1. Participants

The current study has been evaluated and approved by the Regional Committee for Medical and Health Research Ethics and the Data Protection Official (Personvernombudet).

Sixty-two individuals agreed to participate in this study. Participants were recruited from local educational, governmental, health care, and sports organizations, and public poster boards. In order to participate, individuals signed up through the web-based recruitment system which is part of the Memoro platform, or called the project phone. Exclusion criteria were age < 40, previous or current neurological disease and psychiatric or medical conditions that may influence test performance. In addition, participation in the Nord-Trøndelag health survey (HUNT) (Krokstad et al., 2013) was an exclusion criterion as a Memoro battery is already administered in the HUNT study population. No participants had impairment(s) in vision, hearing, or motor function that could affect their performance. Participants received a monetary reward equivalent of $50 USD after completing the retest session.

2.2. Memoro

Memoro is a self-administered web-based neuropsychological test platform developed to measure memory and related cognitive functions in large cohorts using neuropsychological tests that are familiar to clinicians and researchers. Additionally, the system provides the flexibility of including new tests based on current research (Hansen et al., 2015). The Memoro tests are designed to be resistant to low bandwidth, and to handle variations in software and hardware configurations by preloading stimuli and employing cross-browser compatible code. All tests include both written and auditory instructions. No test in the current battery depends on precise millisecond precision. It has been found feasible to detect reaction time effects as small as ~20 ms on web-based tests (Crump, McDonnell, & Gureckis, 2013), but given the large variety of devices and settings in which the tests might be completed we have focused primarily on accuracy and maximum capacity measures. Memoro employs context measures to address some of the concerns of whether the participant completes the tests in a valid manner. At login the participants answer questions regarding sleep, alertness, and what kind of computer they are using, and are asked to evaluate the noise level in the room.

In the current study participants completed the following tests:

Verbal Memory Test

The participant listened to a target list of 16 words chosen from four different semantic categories (furniture, fruits, animals, and means of transportation). After the complete list was presented, the participant was asked to type all words he/she recalled from the list into boxes presented on the computer screen. There were four learning and recall trials for the target list, followed by a distraction list. The distraction list contained 16 words chosen from four semantic categories, of which three overlapped with the first list (furniture, fruits, animals, and body parts). After hearing the distraction list, the participant recalled and typed in the words from this list, similarly to the procedure for the target list. The participant was then asked to recall the words from the target list, i.e. immediate recall. After ~20 min of completing other non-verbal tests, the participant completed the delayed recall of the target list. Performance was scored as number of correctly recalled words in each trial.

Objects in Grid

The test started with a short practice session of dragging and dropping objects on the screen. Subsequently, the participant was instructed to memorize the locations of 18 colored line drawings of various objects within a 6 × 6 grid in 90 s one-trial learning period. After 90 s, all objects were moved to the bottom of the screen and the participant was instructed to drag and drop each object into its original location in the grid, i.e. immediate recall. After completing other non-spatial tests for ~15 min, the participant was once more presented with the empty grid and asked to place all the objects into their correct locations, i.e. delayed recall. All objects must be placed in the grid for the test to continue, and the participant was instructed to guess if unable to recall the location of an object. Performance was scored as the number of correctly placed objects.

Letter-Number Sequencing

In each trial of this test, the participant was presented with individual letters and numbers on the screen. Each stimulus in a trial was visible for two seconds. After the last stimulus in a trial had been presented, the participant was asked to recall and type first the numbers in ascending order and then the letters in alphabetical order. The participant started with three practice trials where he/she got feedback on whether the responses were correct.

The test session included up to 14 trials where the number of letters and numbers increased by one every second trial from two to a maximum of eight characters. Performance was scored as the number of correct trials, and the stop criterion was three consecutive wrong trials.

Processing Speed

The participant was instructed to judge as fast as possible without making mistakes if pairs of geometrical shapes (block 1, 3, 5) or numbers (block 2, 4, 6) were identical or different by hitting the “F” (different) or “L” (identical) keys on the keyboard. The test started with four example trials (two from each block type) with feedback on the performance. The test did not commence before correct responses were registered. Then the participant completed six blocks, each lasting 30 s. Task difficulty increased with each block as complexity in the geometrical shapes increased and the numbers contained increasingly more digits. Performance was scored as total number of correct responses minus number of erroneous responses combined for both numbers and geometrical shapes.

Pattern Separation

The participant was presented with a series of 108 images, one image at a time, and instructed to indicate whether the current image was identical to a previously presented image, similar to a previously presented image with a detail changed, or never previously presented in the series of images (Bakker, Kirwan, Miller, & Stark, 2008). The next image was not presented before getting a response. The participant received a score of the total number of correct decisions. Being able to discriminate between similar stimuli is an important aspect of episodic memory (Yassa & Stark, 2011).

Computer familiarity

The Memoro Short Computer Questionnaire contains six questions. Three assessing computer usage; “Where have you used a computer during the last 6 months?,” alternatives (score): “Home (2), Work (2), Other (1),”“What activities have you done on a computer during the last 6 months?,” alternatives: “Paying bills (1), E-mail (1), Browsing (1), Office (1), Multimedia (1),”“How often do you use a computer?,” alternatives: “Daily (5), Several times a week (4), Once a week (3), More than once a month (2), Less than once a month (1).” Three assessing computer skill; “How comfortable are you in using a computer mouse?,” alternatives: Very uncomfortable (1), Uncomfortable (2), Neither nor (3), Comfortable (4), Very comfortable (5), “How comfortable are you using a computer keyboard?,” same alternatives as previous question, “How comfortable are you using computers on a scale from 1 to 10 with 1 being ‘only experience problems’ and 10 ‘no problems at all’?”. Two sub-scores, usage (max 15 points) and skill (max 20 points) are combined in a total computer familiarity score (max 35 points).

2.3. Procedure

The same procedure was used for the first and second test sessions. Participants completed the Memoro tests at the MR Center, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway. Each participant was seated in a room equipped with a desktop computer with a 17″ screen (resolution 1024 × 768), standard keyboard, mouse, and speakers attached. The tests were completed in Mozilla Firefox (version 15) running on Microsoft Windows XP. A research assistant gave a short introduction to the computer setup, and made sure the sound volume of the speakers was adjusted to the participants' preference. Before the cognitive tests were administrated, participants answered the context measure questions including sleep quality the previous night, alertness and perceived noise in the test room (see Section 2.2).

In order to avoid possible variations in difficulty between different stimuli sets, we used the same stimuli sets for the first and second test sessions. We chose a minimum test-retest interval of 14 days and strived for the participants to be tested at the same time of day at both sessions. This short test-retest interval would allow us to see how practice effects vary across our cognitive measures in a time window relevant for assessment of effects of experimental manipulation (e.g. memory training) or transient phenomena (e.g. severe hypoglycemia, post-traumatic amnesia).

2.4. Statistical analyses

Raw scores were extracted from the Memoro database into IBM SPSS Statistics for Windows (version 22). The data was quality controlled and scores representing lack of compliance or premature termination of the test were excluded from the analyses. Descriptive statistics are given as frequency, arithmetic mean with standard deviation, or median with inter-quartile range depending on data distribution. Practice effects from test to re-test were assessed with paired samples *t*-tests and effect sizes reported as Cohen's *d* for repeated measurements (Dunlap, Jose, Vaslow, & Burke, 1996; Dunst, Hamby, & Trivette, 2004). Differences in practice effects between sexes were investigated using independent samples *t*-tests. The relationship between practice effects and age and computer familiarity were investigated using Pearson correlation, and Spearman correlation for education and number of days between testing. Test-retest reliability indices were given as Pearson's correlation coefficients and ICCs between raw scores at test and retest. Spearman's *rho* was used to assess reliability of the immediate and delayed recall measures of the Verbal Memory Test since a ceiling effect was present in these measures.

P-values ≤ 0.05 (two-tailed) after application of the Bonferroni-Holm (Holm, 1979) method for correction for multiple comparisons were considered statistically significant.

2.4.1. Intraclass correlation

There are several ICC families and nomenclatures in use. For a thorough discussion of choice and description of intraclass correlation, see Weir (2005). The most common are those of Shrout and Fleiss (1979) and McGraw and Wong (1996). The four different ICCs that are used in this study are presented in Table 1.

Our primary test-retest measure is ICC(2,1), which is equivalent to ICC(A,1). This measure is an estimate of reliability based on a two-way random effects model of variance, examining the absolute agreement between test sessions and estimating the reliability of a single assessment. We arrived at this particular ICC measure based on our responses to topics raised in McGraw and Wong (1996). They are presented below:

1. Is a one-way or a two-way analysis of variance appropriate?
   Since each included participant has a test score from each time point included in the analysis, i.e. the usual situation for test-retest studies, the recommended analysis of choice is a two-way analysis of variance (Weir, 2005).
2. Can effects due to trials/time points be ignored in the reliability index?
   This addresses whether systematic error such as practice effects or fatigue effects due to multiple administrations should be included (absolute agreement type analysis), or ignored (consistency type analysis). It has both been argued that practice effects are a natural phenomenon and not a defect of the procedure (Rousson et al., 2002), and that practice effects represent error which should be addressed in the analysis (Weir, 2005). In

Table 1
Nomenclature and formula of ICCs used in this study.

Shrout & Fleiss	McGraw & Wong	Formula	Short description
2, 1	A, 1	$\frac{MS_p - MS_e}{MS_p + (k-1)MS_e + \frac{k(MS_t - MS_e)}{n}}$	Two-way random effects model, agreement, single measures
2, k	A, k	$\frac{MS_p - MS_e}{MS_p + \frac{(MS_t - MS_e)}{n}}$	Two-way random effects model, agreement, average measures
3, 1	C, 1	$\frac{MS_p - MS_e}{MS_p + (k-1)MS_e}$	Two-way mixed effects model, consistency, single measures
3, k	C, k	$\frac{MS_p - MS_e}{MS_p}$	Two-way mixed effects model, consistency, average measures

In Shrout and Fleiss (1979) the abbreviations are, 2 = two-way random effects model, 3 = two-way mixed effects model. In McGraw and Wong (1996), A = Degree of absolute agreement among measurements, C = Degree of consistency among measurements. In the formula column the following abbreviations were used: MS_p = participants' mean square, MS_e = error mean square, MS_t = trials (test sessions) mean square, k = number of trials, n = number of participants. All measures can be derived from the summary table of a repeated measures ANOVA.

the current study, we chose to present both analyses as both are used in the literature. However, as our main measure of reliability we chose the recommended absolute agreement type of analysis (Weir, 2005).

3. Random or Mixed effects for trials/time points?
   This decision has no effect on the calculation and resulting estimate of the ICC, but relates more to its interpretation. With a random effects model, both participants and trials are considered as sampled from a larger pool of participants and trials which in turn the results can be generalized to. In a mixed effects model, participants are still considered sampled from a random pool of participants while trials are fixed (i.e. considered the only trials of interest) and subsequently the results may not be generalized beyond bounds of the study. The current study use both random and mixed effects models. A more detailed discussion of this topic is found elsewhere (McGraw & Wong, 1996; Weir, 2005).
4. What is the unit of reliability?
   This is a question of whether we want to estimate the reliability of a score based on one administration of a test (single measures) or on the combination of scores derived from multiple administrations (average measures). Average measures will always produce larger coefficients as the impact of measurement error is reduced. The appropriate measure to report depends on whether reliability should be estimated for a single administration of the test (the most typical situation), or should be estimated for the combination (an average) of multiple administrations. We use the single measures estimate as our primary measure because participants need only complete the Memoro tests once in order to obtain a score. The average measure is included for all tests in our results section to illustrate how it differs from single measures and for comparison of the present results to other studies that presented average measures.

Summarized, the primary reliability measure was based on a two-way random effects model of variance with an absolute agreement definition, reporting single measures. This is the ICC(2,1) in the Shrout and Fleiss nomenclature, ICC(A,1) in the McGraw and Wong nomenclature.

Frequently used statistical software such as STATA and R (psych library) present results using both Shrout and Fleiss (1979) and McGraw and Wong (1996) nomenclatures, and by default present all ICC variants used in this study. SPSS uses the McGraw and Wong (1996) nomenclature, with the two-way mixed effects model and degree of consistency (i.e., ICC(3,1) and ICC(3,k)) as the default output.

3. Results

3.1. Sample characteristics

One participant of the original 62 did not show up for the re-test session and was unavailable for setting up a new appointment. Sixty-one participants (31 females) with *Mean (SD)* age of 53.3 (7.3) years completed the test and re-test sessions with a *Median* (*IQR*) test-retest interval of 14 (14–19) days. Level of education among the participants was completed secondary school (3.3%), completed high school (18%), completed technical college (3.3%), completed three or fewer years of college or university education (23%), and more than three years of college or university education (52.5%). The computer familiarity score from the Memoro Short Computer Questionnaire was *Mean* (*SD*) 26.77 (3.59) out of 35, indicating extensive familiarity with computers.

3.2. Practice effects

Significant differences in raw scores between the two time points were found on all measures except the computer familiarity questionnaire (Table 2). The performance on the second time point was better than the first time demonstrating practice effects as expected. The Verbal Memory Test showed pronounced practice effects, and had ceiling effects. A large practice effect was also observed for the Images in Grid spatial memory task, but this did not result in any ceiling effects.

There were no significant differences in practice effects between the sexes on any measure ($p > 0.05$). There were no significant (all $p > 0.05$) correlations between practice effects and levels of education or computer familiarity.

3.3. Test-retest reliability

Processing Speed had the highest estimated reliability among the cognitive measures with ICC(2,1) = 0.74 (Table 3). The memory related measures had lower ICC(2,1) coefficients ranging between 0.55 and 0.67 (Table 3). Measures with larger practice effects generally have lower reliability ICC(2,1) coefficients. If practice effects are not considered a problem, we can use ICC(3,1). ICC(3,1) was 0.80 for Processing Speed, the memory scores ranged from 0.65 to 0. 72 with Objects in Grid late recall being the only test below 0.70 with an estimated ICC(3,1) of 0.65.

Finally, we estimated the reliability of combination of the scores from two test administrations, i.e. average ICC measures, ICC(2,2) and ICC(3.2). For the cognitive measures we found that ICC(2,2) ranged between 0.71 and 0.85 and ICC(3,2) between 0.79 and 0.89. The measure of Computer familiarity was estimated to 0.90 for both ICC(2,2) and ICC(3,2).

Table 2
Raw scores at time point 1 and 2 and size of practice effects.

Measure	Time 1	Time 2	*p*	*d*
Computer familiarity	26.98 (3.30)	27.41 (3.28)	0.10	0.13
Letter-Number Sequencing	9.33 (2.50)	10.15 (2.81)	< 0.01	0.34
Pattern Separation	95.41 (3.98)	96.90 (5.03)	< 0.01	0.32
Processing Speed	59.80 (14.55)	65.74 (13.92)	< 0.01	0.42
Objects in Grid				
Immediate recall	9.72 (4.01)	12.11 (3.99)	< 0.01	0.60
Delayed recall	8.4 (4.05)	11.05 (4.34)	< 0.01	0.61
Verbal Memory Test				
Distraction trial	8.9 (2.59)	10.22 (3.07)	< 0.01	0.45
Immediate recall[a]	15 [12–16]	15 [14–16]	< 0.01	0.40
Delayed recall[a]	14 [12–16]	15 [14–16]	< 0.01	0.43

Values given as mean (*SD*) if not stated otherwise.

[a] Scores are negatively skewed with ceiling effects. Central tendency reported as median and inter-quartile range. Statistical difference assessed with Wilcoxon's Signed Rank Test and effect size given as Z/sqrt(n).

3.4. Missing data

The scores of seven participants were invalid on the Letter-Number Sequencing test because they swapped the order of the digits and the letters. Scores from four participants were excluded on the second administration of Objects in Grid due to suspicion of low effort.

4. Discussion

This initial reliability study of Memoro based on the same stimuli set and re-testing after two weeks, demonstrated good test-retest reliability, with absolute agreement based reliability estimates attenuated by practice effects. We also showed how choice of ICC method greatly affected the ICC coefficients, particularly the difference between single measures and average measures ICCs. This highlights the impact of choice of ICC measure.

The measure of processing speed had the highest estimated reliability while the different memory measures had lower reliability estimates. This pattern of speed related tasks showing high reliability and memory measures showing lower reliability is in agreement with research on both traditional paper-and-pencil (Calamia, Markon, & Tranel, 2013), computerized tests (Gualtieri & Johnson, 2006), and web-based (Erlanger et al., 2002) neuropsychological tests.

The Verbal Memory Test had a relatively high reliability despite a ceiling effect. The ceiling effect was present on both the immediate and delayed recall measures resulting in a restricted range. This was not found during piloting or in our previous study in older participants (Hansen et al., 2015). It is not uncommon that memory tests suffer from a restricted range of performance (Strauss, Sherman, & Spreen, 2006). However, in most cases it is reduced memory rather than extraordinary good memory which is of clinical interest. However, ceiling effects implies reduced sensitivity. Without modifications, the current version of the test may be suboptimal for certain groups of individuals, for instance if examining the full range of memory performance. Spearman correlations of 0.70 and 0.67 showed that the Verbal Memory Test has adequate reliability when disregarding practice effects. Similarly, the distraction trial which was not as affected by practice effects or had ceiling effects showed an ICC(2,1) of 0.65 and ICC(3,1) of 0.71, indicating adequate reliability. There are very few web-based verbal memory tests which present the words by sound (i.e. read aloud), and include both immediate and delayed recall trials. The most relevant comparison would be to the CVLT-II (Woods, Delis, Scott, Kramer, & Holdnack, 2006) which is administered in a traditional pen-and-paper format by a trained administrator and found to have estimated reliability with Spearman's *rho* of 0.80 on immediate recall, 0.83 on delayed recall and 0.56 on the distraction trial in a similar sample to ours. This indicates that, despite a ceiling effect, our verbal memory test had on average comparable reliability to the traditional pen-and-paper test.

Objects in Grid had the lowest estimated reliability, where the Objects in Grid delayed recall trial had an ICC(2,1) of 0.55 and ICC(3,1) of 0.65. This test showed the largest practice effects of all tests, and consequently the largest difference between absolute agreement and consistency ICC measures. This illustrates how systematic error affects the estimation of reliability as measured with ICC(2,1). In the Objects in Grid test, participants were to remember the location of 18 line drawings presented only once for a 90 s period. Although we did not observe any flooring effects or restricted range on this test, it may have been perceived as very difficult by the participants due to its one-trial learning design leading to reduced effort. In contrast to the Verbal Memory Test, the Objects in Grid test completion depended on all objects being placed onto the grid, and thus forced the participants to guess. This may introduce additional variance further lowering reliability. Still, the reliability for Objects in Grid is equally good as that reported in previously published test-retest data on similar web-based memory tests (Elbin, Schatz, & Covassin, 2011; Erlanger et al., 2002). The Memory score of the Cognitive Stability Index (CSI), including a test

Table 3
Estimated test-retest correlation coefficients for the Memoro measures in this study.

Measure	*r*	Agreement		Consistency	
		Single ICC(2,1)[a]	Average ICC(2,2)	Single ICC(3,1)	Average ICC(3,2)
Computer familiarity	0.82	0.82	0.90	0.83	0.90
Letter-Number Seq.	0.71	0.67	0.81	0.71	0.83
Pattern Separation	0.72	0.66	0.80	0.70	0.82
Processing Speed	0.80	0.74	0.85	0.80	0.89
Objects in Grid					
Immediate recall	0.72	0.61	0.76	0.72	0.84
Delayed recall	0.65	0.55	0.71	0.65	0.79
Verbal Memory Test					
Distraction trial	0.72	0.65	0.79	0.71	0.83
Immediate recall[b]	0.70				
Delayed recall[b]	0.67				

All correlation coefficients were statistically significant ($p < 0.05$).

ICC(2,1) = absolute agreement, single measures; ICC(2,2) = absolute agreement, average measures; ICC(3,1) = consistency, single measures; ICC(3,2) = consistency, average measures.

[a] Primary ICC test-retest measure.

[b] Spearman's *rho* used on data with skew and ceiling effect.

similar to the Objects in Grid test, had a Pearson test-retest correlation coefficient of 0.68 (Erlanger et al., 2002). We found Pearson correlations of 0.72 and 0.65 for immediate and delayed recall respectively. Another study using the online version of ImPACT (Elbin et al., 2011) reported a Pearson correlation of 0.55 and an ICC(3,2) of 0.70 on the Visual Memory composite that includes two tests which combined is similar to our Objects in Grid task. We found ICC(3,2) of 0.84 and 0.79 for immediate and delayed recall, respectively. However, ICC(3,2) is a measure of the estimated reliability of a combined score from two administrations of the test and not an estimate of the reliability of a single administration. Taken together, even though the Objects in Grid test was the test with the lowest test-retest reliability estimate, it showed similar reliability estimates as comparable tests from other web-based test batteries.

The Pattern Separation and Letter-Number Sequencing tasks showed moderate reliability. The Pattern Separation task was mildly affected by practice effects compared to the other memory tests in this study. The CogState Brief Battery has implemented a pattern separation paradigm in its One Card Learning Task. In a sample of community-dwelling older people they estimated test-retest reliability of each measure across five visits using intraclass correlations and found an ICC of 0.91 for this task (Darby et al., 2014). Our Pattern Separation task had an ICC(2,1) of 0.66, still, our average ICC(2,2) was 0.80 and could have been even higher if more time points were included. Indeed, using the consistency measure, ignoring the practice effects, the average ICC(3,2) value for Pattern Separation was 0.82. Taken together, the reliability estimates for our Pattern Separation task appeared to be somewhat lower but comparable to a similar task used previously. Still, use of unspecified ICC analyses makes it difficult to compare between studies and know which reliability has been estimated.

The Letter-Number Sequencing test was to a limited extent affected by practice effects compared to the other tests in this study. This finding is in line with previously reported reliability on this type of test. Using a pen-and-paper format (Lemay, Bédard, Rouleau, & Tremblay, 2004) had a test-retest interval of 14 days in a sample of adults and seniors and found Pearson correlation of 0.75 and an ICC with an absolute agreement definition of 0.73. Our Letter-Number Sequencing test had a lower single measure absolute agreement ICC(2,1) of 0.67, and a higher average measure absolute agreement ICC(2,2) of 0.81 The authors (Lemay et al., 2004) did not state whether they report single or average measure ICC reliability estimates.

The computer familiarity measure included in Memoro was found to be highly reliable. Computer familiarity has previously been shown to affect test scores (Iverson, Brooks, Ashton, Johnson, & Gualtieri, 2009) and be related to general cognitive function (Tun & Lachman, 2010). In Memoro, we have previously reported an association between computer familiarity and cognitive measures, although affected by age and education, that was most pronounced for our processing speed task (Hansen et al., 2015). In the present study we did not find computer familiarity to be related to practice effects, probably due to extensive computer familiarity across the entire sample. Based on these combined observations, information regarding computer familiarity should be obtained when using web-based neuropsychological testing in order to investigate whether it should be controlled for or not.

Overall, the findings across the different tests indicated that the measurement (random) error of the Memoro tests is acceptably low. Test-retest reliability estimated with an absolute agreement definition might be higher with a longer retest interval than two weeks, decreasing the practice effects. However, practice effects should not represent a problem if not resulting in ceiling effects and relative group differences are the measures of interest.

There are important discrepancies in the use and reporting of results from intraclass correlation in reliability studies. In our own data, Pearson's *r* coefficients, which do not take systematic error in to account, ranged between 0.65 and 0.80. ICCs, estimating the reliability of a score based on a single administration of each test, ranged between 0.55 and 0.74 if taking systematic error into account, and from 0.65 to 0.80 if not. ICCs estimating the reliability of a score based on the combined results from two administrations of each test ranged between 0.71 and 0.85 if taking systematic error into account and from 0.79 to 0.89 if not. We have observed that several studies have reported average measures ICCs without making it explicit that this is not the estimated reliability of a single administration of the battery or test, and neither does it unequivocally represent a measure for the tests' stability over time. ICC results are often used in conjunction with clinically accepted cutoffs or suggested thresholds; hence it is important to report ICC coefficients which actually estimate the relevant reliability relating directly to the use of the test in the clinic or in research. Moreover, the inconsistent ICC use and reporting becomes problematic in literature reviews, because different ICC variants are included under the generic heading "ICC" without additional comments. To sum up, ICC is not a unitary measure and it is therefore important to carefully choose the appropriate ICC and specify it in detail when reporting test-retest results.

The current study has limitations worth noting. First, several participants misunderstood the instructions on the Letter-Number Sequencing task and swapped the position of the numbers and letters resulting in erroneous responses in spite of recalling the stimuli correctly. To reduce the likelihood of this problem in the future we will include a notice on the response form reminding the participant of the instruction. This together with missing data on the Objects in Grid test illustrate some of the vulnerabilities with self-administered tests as the administrator is not present to detect if an instruction is misunderstood or the participant is not performing at full effort, and take appropriate steps to rectify performance. Second, due to a combination of high-performing participants and the choice of using a short test-retest interval we observed significant practice effects on all cognitive measures and ceiling effects on the Verbal Memory Test. In the future we need to consider reducing the number of learning trials or increasing the number of items in order to add more difficulty to the Verbal Memory Test, especially if using this test in groups we suspect to have high initial scores. Despite that, it has been argued that practice effects are a natural phenomenon (Rousson et al., 2002) and lack of it, or difference in practice effects between groups can provide meaningful information on learning ability (Chelune, Naugle, Lüders, & Sedlak, 1993).

5. Conclusions

The results of the present study showed that the selected Memoro tests have test-retest reliability similar to traditional, computerized and web-based test batteries used in the clinic and in research. Three weeks or shorter test-retest intervals are not recommended for the Memoro memory measures unless using alternative stimuli sets, due to practice effects reducing the stability of performance, especially if absolute differences are of interest. Explicit specification of which ICC variant is used and whether single or average measures are reported, and providing a rationale behind this choice is very important. Our results showed that the discrepancy between the different measures can be large, thus significantly influencing the interpretation of the test scores and how reliable a test is perceived to be. While Memoro shows reliability coefficients similar to other batteries, a higher level of reliability is the ideal, especially for clinical use. In the future development and quality assurance of Memoro it will be important to investigate and address technical and human factors that may have contributed to measurement error and reduced reliability.

References

Bakker, A., Kirwan, C. B., Miller, M., & Stark, C. E. L. (2008). Pattern separation in the human hippocampal CA3 and dentate gyrus. *Science (New York, NY), 319*(5870), 1640–1642. http://dx.doi.org/10.1126/science.1152882.

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. http://dx.doi.org/10.1080/13854046.2012.663001.

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

Bruce, J., Echemendia, R., Meeuwisse, W., Comper, P., & Sisco, A. (2014). 1 year test-retest reliability of ImPACT in professional ice hockey players. *The Clinical Neuropsychologist, 28*(1), 14–25. http://dx.doi.org/10.1080/13854046.2013.866272.

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

Chelune, G. J., Naugle, R. I., Lüders, H., & Sedlak, J. (1993). Individual change after epilepsy surgery: practice effects and base-rate information. *Neuropsychology, 7*(1), 41–52. http://dx.doi.org/10.1037/0894-4105.7.1.41.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One, 8*(3), e57410. http://dx.doi.org/10.1371/journal.pone.0057410.

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. http://dx.doi.org/10.1016/j.chb.2013.08.009.

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

Dunlap, W. P., Jose, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1*(2), 170–177. http://dx.doi.org/10.1037/1082-989X.1.2.170.

Dunst, C. J., Hamby, D. W., & Trivette, C. M. (2004). Guidelines for calculating effect sizes for practice-based research syntheses. *Centerscope, 3*(1), 1–10.

Elbin, R. J., Schatz, P., & Covassin, T. (2011). One-year test-retest reliability of the online version of ImPACT in high school athletes. *The American Journal of Sports Medicine, 39*(11), 2319–2324. http://dx.doi.org/10.1177/0363546511417173.

Erlanger, D., Feldman, D., Kutner, K., Kaushik, T., Kroger, H., Festa, J., et al. (2003). Development and validation of a web-based neuropsychological test protocol for sports-related concussion. *Archives of Clinical Neuropsychology, 18*, 293–316.

Erlanger, D. M., Kaushik, T., Broshek, D., Freeman, J., Feldman, D., & Festa, J. (2002). Development and validation of a web-based screening tool for monitoring cognitive status. *The Journal of Head Trauma Rehabilitation, 17*(5), 458–476. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12802255.

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

Gualtieri, C. T., & Johnson, L. G. (2006). Reliability and validity of a computerized neurocognitive test battery, CNS Vital Signs. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists, 21*(7), 623–643. http://dx.doi.org/10.1016/j.acn.2006.05.007.

Hansen, T. I., Haferstrom, E. C. D., Brunner, J. F., Lehn, H., & Håberg, 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*, 1–14. http://dx.doi.org/10.1080/13803395.2015.1038220.

Haworth, C., & Harlaar, N. (2007). Internet cognitive testing of large samples needed in genetic research. *Twin Research and Human Genetics, 10*(4), 554–563. Retrieved from http://journals.cambridge.org/abstract_S1832427400008148.

Holm, S. (1979). A simple sequential rejective method procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

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. http://dx.doi.org/10.1080/13803390802372125.

Krebs, D. (1986). Declare your ICC type. *Physical Therapy, 66*(1431). Retrieved from http://ptjournal.apta.org/content/66/9/1431.1.short.

Krokstad, S., Langhammer, A., Hveem, K., Holmen, T. L., Midthjell, K., Stene, T. R., et al. (2013). Cohort Profile: the HUNT Study, Norway. *International Journal of Epidemiology, 42*(4), 968–977. http://dx.doi.org/10.1093/ije/dys095.

Lemay, S., Bédard, M.-A., Rouleau, I., & Tremblay, P.-L. 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. http://dx.doi.org/10.1080/13854040490501718.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46. http://dx.doi.org/10.1037//1082-989X.1.1.30.

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. http://dx.doi.org/10.1016/j.cmpb.2005.06.007.

Morley, J. E., Morris, J. C., Berg-Weger, M., Borson, S., Carpenter, B. D., Del Campo, N., et al. (2015). Brain health: the importance of recognizing cognitive impairment: an IAGG consensus conference. *Journal of the American Medical Directors Association, 16*(9), 731–739. http://dx.doi.org/10.1016/j.jamda.2015.06.017.

Resch, J. E., McCrea, M. A., & Cullum, C. M. (2013). Computerized neurocognitive testing in the management of sport-related concussion: an update. *Neuropsychology Review, 23*(4), 335–349. http://dx.doi.org/10.1007/s11065-013-9242-5.

Rousson, V., Gasser, T., & Seifert, B. (2002). Assessing intrarater, interrater and test-retest reliability of continuous measurements. *Statistics in Medicine, 21*(22), 3431–3446. http://dx.doi.org/10.1002/sim.1253.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18839484.

Silverstein, S. M., Berten, S., Olson, P., Paul, R., Willams, L. M., Cooper, N., et al. (2007). Development and validation of a World-Wide-Web-based neurocognitive assessment battery: WebNeuro. *Behavior Research Methods, 39*(4), 940–949. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18183911.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A Compendium of neuropsychological tests: administration, norms, and commentary* (3rd ed.). New York: 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. http://dx.doi.org/10.1002/gps.3993.

Tun, P. A., & Lachman, M. E. (2010). The association between computer use and cognition across adulthood: use it so you won't lose it? *Psychology and Aging, 25*(3), 560–568. http://dx.doi.org/10.1037/a0019543.

Vaz, S., Falkmer, T., Passmore, A. E., Parsons, R., & Andreou, P. (2013). The case for using the repeatability coefficient when calculating test-retest reliability. *PLoS One, 8*(9), e73990. http://dx.doi.org/10.1371/journal.pone.0073990.

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.

Woods, S. P., Delis, D. C., Scott, J. C., Kramer, J. H., & Holdnack, J. A. (2006). The California verbal learning test–second edition: test-retest reliability, practice effects, and reliable change indices for the standard and alternate forms. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists, 21*(5), 413–420. http://dx.doi.org/10.1016/j.acn.2006.06.002.

Yassa, M. A., & Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends in Neurosciences, 34*(10), 515–525. http://dx.doi.org/10.1016/j.tins.2011.06.006.

Zygouris, S., & Tsolaki, M. (2014). Computerized cognitive testing for older adults: a review. *American Journal of Alzheimer's Disease and Other Dementias*. http://dx.doi.org/10.1177/1533317514522852.