Stability of Cognitive and Attentional Computerized Tests • e068007
DOI: https://doi.org/10.14195/1647-8606_68_7

Assessing the Stability of Cognitive and Attentional Computerized Tests: A Test-Retest Reliability Study

Aranza Lira-Delcore[1], Jose L. Tapia[2] and Jon Andoni Duñabeitia[3]

Abstract

This study evaluates the test-retest reliability of four computerized cognitive tests designed to measure selective attention, sustained attention, visual search, and visual scanning. The sample consisted of 39 young adults who completed computerized versions of the Trail Making Test, Visual Search Test, Sustained Attention to Response Test, and Selective Attention Test developed by CogniFit Inc. across three sessions. The study included two comparisons: test-retest on different days and immediate test-retest on the same day. Results indicated consistent outcomes across sessions for all tests, except for certain variables where immediate repetition likely induced learning effects and for variables not crucial for the task. These findings support the reliability of the tests over short intervals, consistent with prior research on cognitive assessments. The study highlights their utility as reliable and accessible tools for broad application. Future research should examine longer intervals and diverse populations to further validate their reliability.

Keywords: attention, cognitive assessment, test-retest reliability, computerized cognitive tests, CogniFit

[1] Centro de Investigación Nebrija en Cognición (CINC), Universidad Nebrija, Madrid, Spain. Email: alira@nebrija.es. ORCID: https://orcid.org/0009-0008-4311-6034.

[2] Centro de Investigación Nebrija en Cognición (CINC), Universidad Nebrija, Madrid, Spain. Email: jtapia@nebrija.es. ORCID: https://orcid.org/0000-0002-9984-2974.

[3] Centro de Investigación Nebrija en Cognición (CINC), Universidad Nebrija, Madrid, Spain. Email: jdunabeitia@nebrija.es. ORCID: https://orcid.org/0000-0002-3312-8559.

Received on 25-09-2024 and approved for publication on 04-06-2025

Avaliando a Estabilidade de Testes Cognitivos e de Atenção Informatizados: Um Estudo de Fiabilidade Teste-Retestes

Resumo

Este estudo avalia a fiabilidade teste-reteste de quatro testes cognitivos informatizados, concebidos para medir a atenção seletiva, a atenção sustentada, a pesquisa visual e a varredura visual. A amostra consistiu em 39 adultos jovens que completaram as versões informatizadas do Trail Making Test, Visual Search Test, Sustained Attention to Response Test e Selective Attention Test, desenvolvidos pela CogniFit Inc., ao longo de três sessões. O estudo incluiu duas comparações: teste-reteste em dias diferentes e teste-reteste imediato no mesmo dia. Os resultados indicaram resultados consistentes entre as sessões para todos os testes, exceto em certas variáveis, onde a repetição imediata provavelmente induziu efeitos de aprendizagem, e em variáveis não cruciais para a tarefa. Estes achados apoiam a fiabilidade dos testes em intervalos curtos, em consonância com investigações anteriores sobre avaliações cognitivas. O estudo sublinha a utilidade destes testes como ferramentas fiáveis e acessíveis para uma ampla aplicação. Pesquisas futuras deverão examinar intervalos mais longos e populações diversificadas para validar ainda mais a sua fiabilidade.

Palavras-chave: atenção, avaliação cognitiva, fiabilidade teste-reteste, testes cognitivos informatizados, CogniFit

INTRODUCTION

In the field of neuropsychology, cognitive science, and clinical practice, precise and consistent evaluation of cognitive functions is fundamental for diagnosing and monitoring individuals' cognitive development. It is widely recognized that cognitive tests aim primarily to detect impairments or patterns of cognitive deficits, whether linked to brain damage or not, by comparing the participant's test score with their reference group (Bird et al., 2004; Cullen et al., 2007; Lachman et al., 2014). However, they can also be used in clinical and educational settings to identify individuals' strengths and weaknesses, which facilitates the design and implementation of programs that help them develop optimally and reach their maximum potential (Flanagan & McDonough, 2022). This is of relevance because cognitive abilities (e.g., processing speed, working memory) have been shown to be valid predictors of academic achievement (Conesa & Duñabeitia, 2021; Tikhomirova et al., 2020).

Cognitive assessments differ from traditional intelligence tests in that they evaluate different areas of an individual's cognitive functioning (such as attention, executive functions, and problem solving) rather than focusing solely on IQ, allowing for a better identification of the domains in which the test-taker excels and those that need improvement (Flanagan et al., 2013; Proust-Lima et al., 2006; Rollè et al., 2019; VandenBos, 2015). However, using traditional (pencil-and-paper) cognitive tests can sometimes be complicated as their administration can be lengthy, tedious, and costly (Lachman et al., 2014; Proust-Lima et al., 2006; Zygouris & Tsolaki, 2015).

One way to facilitate the administration process is through computerized cognitive batteries (Asensio & Duñabeitia, 2023). These instruments differ from pencil-and-paper tests in that they present instructions and items on a computer screen, smartphone, or tablet, and record responses using peripherals or built-in systems (e.g., mouse, touchscreen, touchpad, or keyboard) (Sánchez-Vincitore et al., 2023). They simplify data collection and processing, offer greater control and precision in stimulus presentation and response collection, and record precise timing, offering high chronometric measures, among other advantages. As a result, classic cognitive tasks are increasingly being standardized and digitized, enabling large-scale, cost-efficient testing while improving accessibility and practicality (Sánchez-Vincitore et al., 2023).

Nonetheless, before implementing these instruments, it is important to guarantee the psychometric properties to ensure the success of the tool and to obtain an accurate estimation of cognitive ability (Cullen et al., 2007; Marques-Costa et al., 2018; Morgan et al., 2019; Sürücü & Maslakçi, 2020). These properties include reliability analysis, which refers to the consistency of an instrument (Weir, 2005) for it to perform the same way, as intended, under different conditions (Meeker et al., 2021). In particular, assessing test-retest reliability is a key step in the adaptation process, as it measures the consistency and stability of outcomes achieved when the instrument is administered to the same sample group on separate occasions (Cronbach, 1947; Portney, 2020; Sürücü & Maslakçi, 2020). It indicates the stability of results over time, guaranteeing that the tests used in research, educational, and clinical settings produce stable and repeatable outcomes in the absence of marked learning effects (Bolarinwa, 2015; Portney, 2020). This analysis is necessary as cognitive tests often require multiple administrations at different moments in time (for example, before and after a clinical treatment or as part of a longitudinal study), making it important to assess their stability under repeated measurements (Lee et al., 2021; Noble et al., 2021).

To assess test-retest reliability, participants typically take the same test in two separate moments, assuring the same conditions, and the outcomes are measured using different analyses such as repeated measures ANOVA, Pearson's correlation, or intra-class correlation (Feinkohl et al., 2020; Paiva et al., 2014; Portney, 2020). In the present study, a repeated-measures ANOVA was employed to assess the stability of the measures over time. This method evaluates whether there are significant differences across multiple time points (Correa-Rojas, 2021; da Cunha et al., 2019; Feinkohl et al., 2020; Pautex et al., 2003), detects systematic errors like learning effects (da Cunha et al., 2019; Van Patten et al., 2021), and allows for the calculation of variance components to understand the sources of variability (Correa-Rojas, 2021). Additionally, a two-way intra-class correlation (ICC) analysis was conducted, as it is an accurate analysis to assess the agreement of measures by considering the equality of means and variances. Pearson's correlations were also analyzed to measure the strength of the relationship between the variables in three different moments. However, unlike the ICC, the Pearson coefficient does not account for systematic differences.

This study used CogniFit (CogniFit Inc., San Francisco, US), a digital platform for cognitive evaluation and training that incorporates a battery of evidence-based psychometric instruments to assess various cognitive and brain functions. These tests are similar to, or are based on, other previously validated instruments widely used in cognitive assessment of attentional processes. Specifically, four attention-related tests from the CogniFit portfolio were selected. In cognitive evaluations of attention, traditional paper-based tests have been extensively used. For example, the Trail Making Test (TMT; Reitan, 1958) assesses attention and task-switching skills, requiring participants to connect digits in sequence. The Visual Search Paradigm (Treisman & Gelade, 1980) and its variants measure the ability to find a target among distractors, reflecting visual search efficiency and selective attention. The Sustained Attention to Response Test (SART; Robertson et al., 1997) evaluates sustained attention and inhibitory control by requiring responses to certain stimuli and withholding responses to infrequent targets. Lastly, tests similar to the d2 Test (Brickenkamp, 1962) assess selective attention and concentration by asking participants to identify target symbols among distractors under time constraints across different trials. In this study, we employed digitalized versions of these tests to measure different aspects of attention, such as selective attention, sustained attention, visual search, and visual scanning, and explored their test-retest reliability with various statistical approaches across three critical test moments: the initial test, a delayed repetition the following day, and an immediate repetition on the second test day.

This study had two main objectives: (1) to evaluate the test-retest reliability of four computerized cognitive tasks (Trail Making Test, Visual Search Test, Sustained Attention to Response Test, and Selective Attention Test) designed to measure different aspects of attention, and (2) to compare the consistency of task performance across two different conditions: delayed repetition (across two days) and immediate repetition (within the same session). Based on previous findings using traditional paper-based assessments, it was hypothesized that the computerized tests would demonstrate acceptable to high test-retest reliability across both delayed and immediate intervals.

METHOD

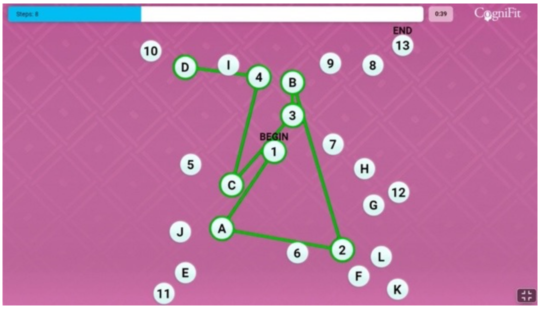
Participants

The sample consisted of 39 participants (27 women; 12 men) who were Spanish residents. Their ages ranged from 18 to 37 years old (*M* = 22.9; *SD* = 3.86). Based on the power analysis conducted beforehand, it was determined that a sample of 38 participants would be sufficient to detect a moderate effect size (*d* = 0.5) with 85% power and a significance level of .05. The inclusion criteria required participants to be between 18 and 40 years of age, without any medical or psychological condition, and to be native Spanish speakers. Participants were recruited from the institution's participant database, attended two sessions each lasting 25 minutes, and received monetary compensation for their participation. The study adhered to the standards of the Declaration of Helsinki, and the procedure was approved by the ethics committee of Universidad Nebrija (approval code UNNE-2022-0017). All participants were informed about the test protocol, given time to ask questions, and asked to sign informed consent forms.

Measures

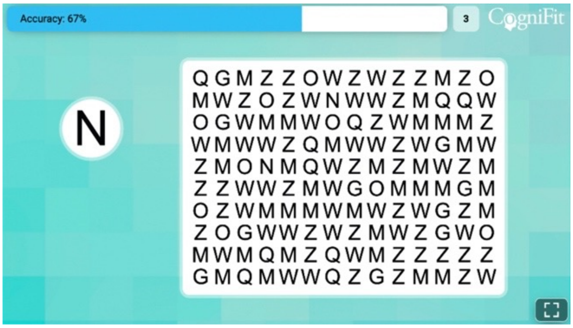
Trail Making Test. This test is a computerized replica of the original test with the same name (Reitan, 1958). It evaluates cognitive processes such as attention, visual search, visual scanning, processing speed, and flexibility, among other executive functions. The task consists of two phases. In the first phase, participants are asked to connect a series of numbers (from 1 to 25) in ascending order by clicking on the corresponding number. This phase measures visual search, visual attention, and processing speed. In the second phase, they are presented with circles containing numbers (from 1 to 13) and letters (from A to L). In this phase, participants must alternate between connecting the numbers and letters (e.g., 1-A-2-B-3-C-...). The second phase also evaluates cognitive flexibility and task-switching skills (see Figure 1). Participants were given 30 seconds to complete each phase. The instrument measures the number of errors and total time in each phase as the main variables of interest.

Figure 1
Example of a trial of the Trail Making Test



Visual Search Test. This task is based on the Visual Search Paradigm (Treisman & Gelade, 1980) and its variants. The test requires the test-taker to identify a target letter within a matrix of letters. The target letter is somehow camouflaged among other letters that serve as distractors, and the participant has 10 seconds per trial to identify it (see Figure 2). In each trial, the target letter appears twice in a 140-letter matrix that, depending on the trial, could include a high or a low proportion of visually similar distractors (e.g., search for N among a high number of W and M competitors). This task measures attention through the cognitive ability of visual scanning and processing speed. It provides information regarding accuracy and response time as the main outcome variables.

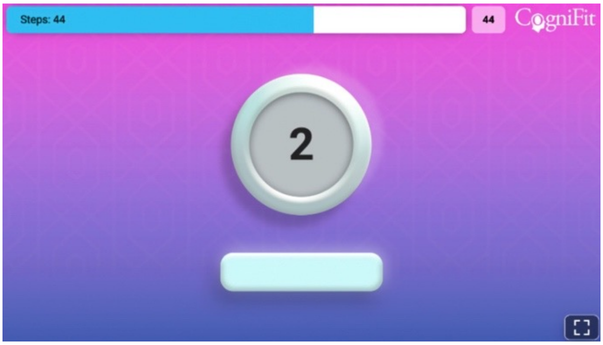
Figure 2
Example of a trial of the Visual Search Test



```
Q G M Z Z O W Z W Z Z M Z O
M W Z O Z W N W W Z M Q Q W
O G W M M W O Q Z W M M M Z
W M W W Z Q M W W Z W G M W
Z M O N M Q W Z M Z M W Z M
Z Z W W Z M W G O M M M G M
O Z W M M M W M W Z W G Z M
Z O G W W Z W Z M W Z G W O
M W M Q M Z Q W M Z Z Z Z Z
G M Q M W W Q Z G Z M M Z W
```

Sustained Attention to Response Test. This test is based on the classic task with the same name (Robertson et al., 1997). The primary goal of this instrument is to measure sustained attention through a fast sequence of numbers. The test-takers are presented with a series of numbers (between 1 and 9), each appearing for 250 milliseconds, and they are told to press the button as quickly as possible when the displayed digit is not a 3 (go trials), withholding response if the number displayed is a 3 (no-go trials) (see Figure 3). The test measures accuracy and response time as the primary outcome variables.

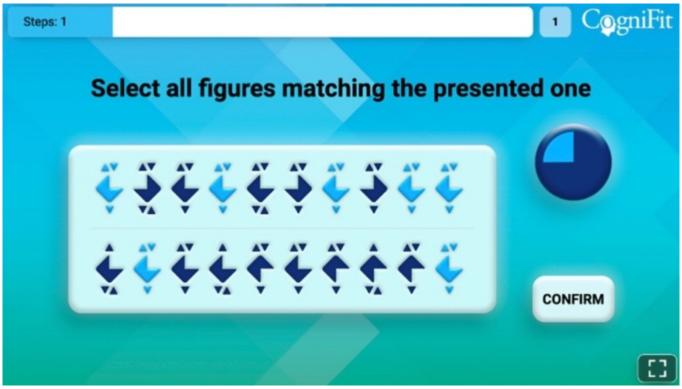
Figure 3
Example of a trial of the Sustained Attention to Response Test



Selective Attention Test. During this task, based on the d2 Test (Brickenkamp, 1962), the test-taker is shown a target stimulus that consists of a figure that can be presented in several variations (four different positions) with some small triangles found either above or below the figure (see Figure 4). The test-taker first learns the visual properties of the target element and then completes five trials of a maximum duration of 30 seconds each, in which the target stimulus appears among a series of distracting stimuli with a high visual similarity to the target one. The test-taker needs to select only the correct stimulus and prevent clicking on the distractors. It provides information regarding accuracy and total completion time as the core resulting variables.

Figure 4

Example of a trial of the Selective Attention Test



Steps: 1

Select all figures matching the presented one

Procedure

A trained evaluator from the research team administered the four computerized cognitive tests in the laboratory. Each test was preceded by a practice session to familiarize participants with the tasks. The assessments were conducted in a quiet and undisturbed environment. The tests were performed on a Lenovo Yoga laptop, featuring an Intel Core i5 8th gen processor, a 1920x1080 32-bit 60 Hz screen, and using an external mouse. The testing procedure spanned two days: the first day (Day 1) consisted of one block of testing, in which participants completed all four tests sequentially, starting with the Trail Making Test, followed by the Visual Search Test, the Sustained Attention to Response Test, and, lastly, the Selective Attention Test. During the second day session (Day 2), which was executed between 24 and 48 hours after Day 1, participants completed two blocks of testing: in the first block (Day 2A), participants completed all four tests sequentially, followed by a 5-minute rest period. After the break, participants repeated the four tests (Day 2B).

Data analysis

For data analysis, the statistical software package Jamovi (v. 2.5.6.0) and the Seolmatrix module (v. 3.9.3) were used. To assess the test-retest reliability of the measurements across test moments, mean scores for each task as measured by the outcome variables at different time points were compared using repeated measures ANOVA. Specifically, comparisons were made between Day 1 to Day 2A (delayed repetition) and Day 2A to Day 2B (immediate repetition). Pearson's correlations were also analyzed to measure the strength of the relationship between the variables in the three different test moments, interpreting the correlation values based on predefined criteria: excellent reliability ($r \geq .80$), good reliability ($.60 \leq r < .80$), fair to moderate reliability ($.40 \leq r < .60$), and low reliability ($r < .40$). Finally, an intra-class correlation (ICC) analysis was conducted, and values were interpreted according to established criteria: low (ICC < .50), moderate (.50 ≤ ICC < .75), good (.75 ≤ ICC < .90), and excellent (ICC ≥ .90) (Liljequist et al., 2019). Analyses were conducted using untransformed data. Statistical significance was established at a p-value of .05.

RESULTS

Descriptive analyses (means and standard deviations) were conducted and are reported in Table 1 for each test moment (Day 1, Day 2A, and Day 2B).

Table 1

Descriptives (means and standard deviations) for each task in each test moment.

Task	Measure	Test Day 1		Re-test Day 2A		Re-test Day 2B		Mean difference	
		M	*SD*	*M*	*SD*	*M*	*SD*	Delayed repetition	Immediate repetition
Trail Making Test (Phase 1)	Number of errors	0.23	0.54	0.23	0.58	0.15	0.43	0.00	0.08
	Total time	20.0	2.93	19.8	3.65	18.8	2.89	0.17	1.04*
Trail Making Test (Phase 2)	Number of errors	0.28	0.65	0.90	4.02	0.31	0.73	-0.62	0.59
	Total time	36.8	7.41	38.8	19.3	32.5	7.20	-1.96	6.24*
Visual Search Test	Accuracy	99.4	4.00	96.2	10.8	96.8	8.47	3.21	-0.64
	Response time	5085	2509	5054	2224	4989	2153	31.40	65.20
Sustained Attention to Response Test	Accuracy	96.5	2.68	96.1	2.43	97.2	1.94	0.40	-1.09*
	Response time	353	58.4	348	71.9	342	72.4	5.57	5.82
Selective Attention Test	Accuracy	97.4	5.51	97.8	3.32	98.1	3.69	-0.41	-0.28
	Total time	74.8	21.4	71.6	18.4	65.1	16.8	3.19	6.55*

Note: Total time is measured in seconds. Response time is measured in milliseconds; Accuracy is measured in percentage of hits. Significant differences at the $p < .05$ level are marked with an asterisk.

For the Trail Making Test, a repeated-measures ANOVA was conducted to specifically examine the effects of test phases (Phase 1 and Phase 2) and time points (Day1, Day2A, and Day2B) on the number of errors and total time, following a 2*3 (i.e., Phase * Time Point) design. The analysis revealed no significant main effect of either Phase ($F(1, 38) = 1.61$, $p = .212$, $\eta^2 = .01$) or Time Point ($F(2, 76) = .88$, $p = .418$, $\eta^2 = .01$) on the number of errors. For total time, there was a significant and large main effect for Phase ($F(1, 38) = 194.79$, $p < .001$, $\eta^2 = .44$) and a significant but small effect for Time Point ($F(2, 76) = 3.94$, $p = .024$, $\eta^2 = .02$). There was no significant interaction effect between the factors for either number of errors (F(2, 76) = .72, $p = .492$, $\eta^2 = .01$) or the total time ($F(2, 76) = 2.62$, $p = .079$, $\eta^2 = .01$). Therefore, both phases of the task were analyzed independently.

Post-hoc analyses (see Table 1 for mean differences) for the Trail Making Test showed no significant differences in the number of errors on Phase 1 when the test-retest is performed with either delayed repetition ($t(38) = .00$, $p = 1.000$) or immediate repetition ($t(38) = .77$, $p = .446$). Similar results were observed in Phase 2 with delayed repetition ($t(38) = -.94$, $p = .354$) and immediate repetition ($t(38) = .89$, $p = .378$), demonstrating consistency in both repetition conditions. For the total completion time, there were no significant differences in delayed repetition in both Phase 1 ($t(38) = -.28$, $p = .784$) and Phase 2 ($t(38) = -.64$, $p = .528$). However, significant differences were observed in immediate repetition in Phase 1 ($t(38) = 2.22$, $p = .032$) and Phase 2 ($t(38) = 2.37$, $p = .023$), suggesting a form of learning or practice effect when the test is completed twice in a row.

For the Visual Search Test, a repeated-measures ANOVA was conducted with a one-factor (Time Point) design with three levels (Day1, Day2A, Day2B) to examine the effects on accuracy and response time. Results showed no significant differences in accuracy ($F(2, 76) = 1.64$; $p = .200$; $\eta^2 = .03$) and response time ($F(2, 76) = .02$; $p = .976$; $\eta^2 = .00$), indicating stability in both measures over time. Post-hoc results (see Table 1) indicate no significant differences in accuracy or response time between the different time points, suggesting consistency in the measurements over time. Specifically, for accuracy, comparisons showed no significant differences in delayed repetition ($t(38) = 1.71$, $p = .096$) and immediate repetition ($t(38) = -.27$, $p = .786$). Similarly, for response time, non-significant differences were observed in the delayed repetition ($t(38) = .07$, $p = .946$) or in the immediate repetition ($t(38) = .16$, $p = .872$). This indicates the stability of the measures of the task over time, showing good test-retest reliability.

Regarding the Sustained Attention to Response Test, a repeated-measures ANOVA was also conducted with a one-factor (Time Point) design with three levels (Day1, Day2A, Day2B) to analyze the effects on accuracy and response time. Results showed significant differences in accuracy ($F(2, 76) = 3.94$; $p = .024$; $\eta^2 = .04$), although the effect size was small. There were no significant differences in response time ($F(2, 76) = 1.02$; $p = .367$; $\eta^2 = .01$). Post-hoc analysis for accuracy (see Table 1) showed there was a non-significant difference between time points in the delayed repetition ($t(38) = 1.00$, $p = .324$), indicating consistent performance between test days. The comparison for immediate repetition revealed a significant difference ($t(38) = -2.71$, $p = .010$), pointing to practice or learning effects when the test is repeated immediately. In terms of response time, comparisons across the same time points did not yield statistically significant differences. Specifically, delayed repetition (t(38) = .66, p = .510) and immediate repetition (t(38) = 1.19, p = .242) showed negligible differences, suggesting stability of the measures over time.

Lastly, for the Selective Attention Test, a repeated-measures ANOVA was conducted with a one-factor (Time Point) design with three levels (Day1, Day2A, Day2B), to examine accuracy and total time. No significant differences were found for accuracy in the Selective Attention Test (F(2, 76) = 1.66; p = .197; η^2 = .01), but significant differences were observed for the total time (F(2, 76) = 7.93; p = <.001; η^2 = .05), though the magnitude of the effect was small. Regarding post-hoc analysis (see Table 1) for accuracy, there were no significant differences in the delayed repetition condition (t(38) = -.996, p = .326), nor in the immediate repetition condition (t(38) = -1.00, p = .324). These findings suggest that accuracy levels remained stable across the testing sessions. For total time, comparisons for delayed repetition showed a non-significant difference (t(38) = 1.43, p = .160), but a significant decrease in total time was found in the immediate repetition condition (t(38) = 3.41, p = .008), suggesting possible practice or learning effects.

As shown in the analysis, descriptive analyses and repeated-measures ANOVA revealed different patterns according to the task. The Selective Attention Test showed high stability in accuracy across each moment, even though total time significantly decreased in immediate repetition. The Sustained Attention to Response Test maintained stable response times, but accuracy significantly improved in the immediate repetition condition only. In the Visual Search Test, both accuracy and response time remained consistent over time, without significant differences. Finally, the Trail Making Test showed significant differences only in total time during immediate repetition, but the number of errors did not show any significant changes in any condition.

Table 2 presents test-retest reliability for each measure across all tasks through Pearson's correlation (r) and Intra-class Correlation Coefficient (ICC). Results showed that the Trail Making Test showed no test-retest reliability of the Number of errors in Phases 1 and 2 in delayed and immediate repetition, estimated by both Pearson's correlation and ICC. Total time in Phase 1 demonstrated good reliability for immediate repetition through Pearson's correlation (r = .62, p = < .001) and ICC (ICC2,1 = .61, F(38, 38) = 4.10, p < .001) and lower but significant reliability for delayed repetition (r = .36, p = .012) (ICC2,1 = .35, F(38, 38) = 2.09, p = .013). Total time in Phase 2 showed moderate reliability for immediate repetition (r = .56, p = < .001) (ICC2,1 = .37, F(38, 38) = 2.15, p = .010) but did not show reliability when analyzing correlations for delayed repetition.

The Visual Search Test did not show significant correlations as estimated by both Pearson's correlation and ICC for accuracy, either in delayed or immediate repetition. Response time exhibited low but significant reliability for both delayed (r = .27, p = .047) (ICC2,1 = .27, F(38, 38) = 1.74, p = .046) and immediate repetition (r = .34, p = .017) (ICC2,1 = .34, F(38, 38) = 2.03, p = .016).

The Sustained Attention to Response Test demonstrated reliability for accuracy across all time points, particularly for delayed repetition, where it showed moderate reliability according to Pearson's correlation (r = .55, p < .001) and ICC (ICC2,1 = .55, F(38, 38) = 3.44, p < .001). Immediate repetition exhibited low but significant reliability (r = .36, p = .012) (ICC2,1 = .35, F(38, 38) = 2.08, p = .013). Response Time showed excellent reliability for immediate repetition (r = .910, < .001) (ICC2,1 = .91, F(38, 38) = 21.30, p < .001). Delayed repetition exhibited good reliability for delayed repetition through Pearson's correlation (r = .70, < .001), but moderate reliability according to ICC criteria (ICC2,1 = .681, F(38, 38) = 5.26, p < .001).

Lastly, the Selective Attention Test total time exhibited very high reliability of accuracy in both delayed (r = .95, < .001) (ICC2,1 = .84, F(38, 38) = 11.50, p < .001) and immediate repetition (r = .88, < .001) (ICC2,1 = .87, F(38, 38) = 14.90, p < .001). Regarding total time, it showed good reliability for delayed repetition (r = .77, < .001) (ICC2,1 = .76, F(38, 38) = 7.24, p < .001), and immediate repetition (r = .66, < .001) (ICC2,1 = .65, F(38, 38) = 4.74, p < .001).

Table 2
Test–retest reliability and effect sizes on four cognitive tests.

Task	Measure	Delayed repetition			Immediate repetition		
		r	ICC$_{2,1}$	F	r	ICC$_{2,1}$	F
Trail Making Test (Phase 1)	Number of errors	-.09	-.09	0.83	.24	.26	1.71
	Total time	-.36*	.35	2.09*	.62***	.61	4.10***
Trail Making Test (Phase 2)	Number of errors	-.04	-.01	0.98	-.06	-.02	0.96
	Total time	.21	.14	1.33	.56***	.37	2.15**
Visual Search Test	Accuracy	-.06	-.04	0.93	-.14	-.14	0.76
	Response time	.27*	.27	1.74*	.34*	.34	2.03*
Sustained Attention to Response Test	Accuracy	.55***	.55	3.44***	.36*	.35	2.08*
	Response time	.70***	.68	5.26***	.91***	.91	21.30***
Selective Attention Test	Accuracy	.95***	.84	11.50***	.88***	.87	14.90***
	Total time	.77***	.76	7.24***	.66***	.65	4.74***

Note: *p < .05, **p < .01, ***p < .001; H_1 is a positive correlation (r).

Overall, consistent test-retest reliability was found for all tasks, although the magnitude varied as a function of the task and the specific measures. The Selective Attention Test showed the most reliable measures, in terms of both accuracy and total time, across immediate and delayed repetition. The Sustained Attention to Response Test also showed good reliability in both accuracy and response time. The Trail Making Test showed moderate reliability in total time, while the number of errors showed low or null reliability. Finally, the Visual Search Test showed significant test-retest consistency for response time.

DISCUSSION

The objective of this study was to assess the test-retest reliability of various cognitive and attentional computerized tests when performed with a delayed repetition and an immediate repetition using repeated measures ANOVAs, Pearson's correlations, and Intra-class correlations. Understanding the consistency of these measures over time is fundamental for their application in both research and practical settings. The findings of the current study indicate that the scores from the four cognitive tasks demonstrated strong stability across multiple time points, particularly when assessments were spaced over a 24–48-hour interval, supporting robust test-retest reliability. Although some practice effects emerged, especially when tasks were repeated with only a short 5-minute interval between assessments, these effects were less pronounced in delayed repetitions, highlighting the overall consistency and reliability of the tasks across time.

Regarding the Trail Making Test, the measurements remained consistent when comparing mean scores of the different time points, except for total time in the immediate repetition condition (Day 2A and Day 2B), which could most likely be explained by practice effects (Chen et al., 2020; Lee et al., 2021; Lee et al., 2023). Pearson's correlation and ICC exhibited reliability mostly for total time, which the original test considers as the primary metric, with the Number of errors serving as a supplementary indicator that does not measure the main variable of interest (Bowie & Harvey, 2006; Reitan, 1958). Therefore, the test variable of interest shows consistency and stability over time, showing that it is a reliable instrument. These results are consistent with previous research that has demonstrated the stability of the Trail Making Test scores in both phases across different moments, whether assessed immediately following each other (Wagner et al., 2011) or with a one-week interval (Woods et al., 2015). Thus, this computerized version of the Trail Making Test exhibits comparable psychometric properties to the original test, indicating consistent results when administered immediately or with intervals of one to two days.

For the Visual Search Test, ANOVAs indicated that the measures used do not vary significantly across the different testing days, which exhibits some stability over time. Additionally, response time showed adequate test-retest reliability when considering Pearson's correlation. These findings align with previous research showing that visual search tasks similar to the Visual Search Test demonstrate good test-retest reliability in both typical individuals and participants with diagnosed impairments (Erez et al., 2009; MacKeben & Fletcher, 2011). The variability in correlations that were not significant could be attributed to differences in cognitive load or fatigue over the testing period. Furthermore, it should be considered that there is a clear ceiling effect in accuracy, as most individuals consistently achieved high scores with very occasional errors. Thus, it is processing speed as a chronometric marker of cognitive processing measured in terms of response time that plays a fundamental role in evaluating the visuospatial abilities that are central to visual search skills (Aul et al., 2023; Murray & Janelle, 2003).

Results for the Sustained Attention to Response Test showed differences circumscribed to the immediate repetition (Day 2A and Day 2B) condition, but not to delayed repetition (Day 1 and Day 2A) which suggests that the test is very consistent over time but that that it is sensitive to practice or learning effects happening in circumstances in which the test is repeated almost immediately. Pearson's correlation and ICC also showed reliability across all time points in both measures, demonstrating that the test is a reliable instrument.

Finally, regarding the Selective Attention Test, accuracy did not show differences between time points, which indicates that the measure that evaluates the ability to select a stimulus among distractors is highly reliable across repetitions. For total time, ANOVA showed differences when comparing time points in the immediate repetition condition (Day 2A with Day 2B), which suggests that this measure is sensitive to practice effects when the test is repeated with a minimal delay between runs. Nonetheless, despite these differences, the remaining analyses showed very good reliability for both measures. Instruments assessing sustained and selective attention, such as the d2 test (which served as the basis for the Selective Attention Test), have previously demonstrated good test-retest reliability across intervals of one week and one month (Lee et al., 2018; Steinborn et al., 2018), in line with the current findings. This consistency suggests that the Sustained Attention to Response Test and the Selective Attention Test align well with other variants of these measures when administered over even shorter time intervals.

Most preceding studies used longer intervals than the ones used in the current experimental setting to assess test-retest reliability, such as days, weeks, or even months, which helps minimize memory effects and provide a more accurate measure of stability over time. The current setting represents a step forward insofar as the interval selected served to assess immediate and short-delay reliability.

However, some limitations should also be considered. One is the relatively small and homogeneous sample, composed of healthy young adults. Although the sample size was justified through a priori power analysis, larger studies including broader demographic profiles (e.g., older adults, individuals with varying educational or cultural backgrounds) are needed to increase the generalizability of the findings.

Additionally, all participants were healthy individuals, which limits the applicability of the results to clinical populations. Future research should examine the test-retest reliability of these instruments in populations for whom attentional and executive functioning assessments are especially relevant, such as individuals with ADHD, traumatic brain injury, or early-stage neurodegenerative conditions. Another limitation is in regard to the presence of a ceiling accuracy effect in most tasks, where the mean percentages of hits consistently exceeded 95% or were very close to this threshold. This ceiling effect reduces the variability in the data, which can hide certain differences and limit the ability to detect subtle variations in performance in more demanding scenarios, or with pathological populations. In this line, it is worth noting that the study sample might not be representative of all demographic groups, which could limit the generalizability of the results. Therefore, future research should explore test-retest reliability over longer periods of time to better understand the stability of the results in the long term. Additionally, examining the applicability and reliability of these tests in diverse populations, including various age groups and clinical populations, would provide a more comprehensive validation of their utility.

Given its demonstrated reliability and user-friendly design, we conclude that the current tests can be effectively used in both research, educational, and clinical settings. An additional advantage of these computerized tests is that they include critical variations in the precise materials (i.e., target items, distractor items) used every time the task is run, while still measuring the same cognitive constructs with an identical procedure. This approach minimizes the replication of exact items in the tasks and mitigates the potential appearance of the practice effect, as seen in the results. By enabling the dynamic generation of new trials on each administration, these tests offer a high degree of variability within a controlled structure, supporting repeated assessments without compromising construct validity. Moreover, the convenience of computerized administration allows for widespread use, including remote assessments, enhancing accessibility for a broader range of users. Therefore, these tests stand as quick and easy-to-use tools for getting an accurate picture of various cognitive abilities, while maintaining the psychometric properties of the original tests or similar versions that measure the same constructs.

REFERENCES

Asensio, D., & Duñabeitia, J. A. (2023). The necessary, albeit belated, transition to computerized cognitive assessment. *Frontiers in Psychology, 14*, 1160554. https://doi.org/10.3389/fpsyg.2023.1160554

Aul, C., Brau, J. M., Sugarman, A., DeGutis, J. M., Germine, L. T., Esterman, M., McGlinchey, R. E., & Fortenbaugh, F. C. (2023). The functional relevance of visuospatial processing speed across the lifespan. *Cognitive Research: Principles and Implications, 8*(1), 51. https://doi.org/10.1186/s41235-023-00504-y

Bird, C. M., Papadopoulou, K., Ricciardelli, P., Rossor, M. N., & Cipolotti, L. (2004). Monitoring cognitive changes: Psychometric properties of six cognitive tests. *British Journal of Clinical Psychology, 43*(2), 197–210. https://doi.org/10.1348/014466504323088051

Bolarinwa, O. (2015). Principles and methods of validity and reliability testing of questionnaires used in social and health science researches. *Nigerian Postgraduate Medical Journal, 22*(4), 195. https://doi.org/10.4103/1117-1936.173959

Bowie, C. R., & Harvey, P. D. (2006). Administration and interpretation of the Trail Making Test. *Nature Protocols, 1*(5), 2277–2281. https://doi.org/10.1038/nprot.2006.390

Brickenkamp, R. (1962). Test d2: Attention stress test.

Chen, K.-W., Lin, G.-H., Chen, N.-C., Wang, J.-K., & Hsieh, C.-L. (2020). Practice Effects and Test–Retest Reliability of the Continuous Performance Test, Identical Pairs Version in Patients with Schizophrenia over Four Serial Assessments. *Archives of Clinical Neuropsychology, 35*(5), 545–552. https://doi.org/10.1093/arclin/acaa004

Conesa, P. J., & Duñabeitia, J. A. (2021). Effects of computer-based training on children's executive functions and academic achievement. *The Journal of Educational Research, 114*(6), 562–571. https://doi.org/10.1080/00220671.2021.1998881

Correa-Rojas, J. (2021). Coeficiente de Correlación Intraclase: Aplicaciones para estimar la estabilidad temporal de un instrumento de medida. *Ciencias Psicológicas, 15*(2), e-2318. https://doi.org/10.22235/cp.v15i2.2318

Cronbach, L. J. (1947). Test "Reliability": Its Meaning and Determination. *Psychometrika, 12*(1), 1–16. https://doi.org/10.1007/BF02289289

Cullen, B., O'Neill, B., Evans, J. J., Coen, R. F., & Lawlor, B. A. (2007). A review of screening tests for cognitive impairment. *Journal of Neurology, Neurosurgery &amp; Psychiatry, 78*(8), 790–799. https://doi.org/10.1136/jnnp.2006.095414

da Cunha, D., Rocha, C., Leite, D. C., Schoenfeld, B. J., & Prestes, J. (2019). New insights for statistical analysis of blood pressure response to resistance training in elderly hypertensive women. *Journal of Physical Education, 30*, e3025. https://doi.org/10.4025/jphyseduc.v30i1.3025

Erez, A. B.-H., Katz, N., Ring, H., & Soroker, N. (2009). Assessment of spatial neglect using computerised feature and conjunction visual search tasks. *Neuropsychological Rehabilitation, 19*(5), 677–695. https://doi.org/10.1080/09602010802711160

Feinkohl, I., Borchers, F., Burkhardt, S., Krampe, H., Kraft, A., Speidel, S., Kant, I. M. J., Van Montfort, S. J. T., Aarts, E., Kruppa, J., Slooter, A., Winterer, G., Pischon, T., & Spies, C. (2020). Stability of neuropsychological test performance in older adults serving as normative controls for a study on postoperative cognitive dysfunction. *BMC Research Notes, 13*(1), 55. https://doi.org/10.1186/s13104-020-4919-3

Flanagan, D. P., & McDonough, E. M. (2022). *Contemporary Intellectual Assessment: Theories, Tests, and Issues*. Guilford Publications.

Flanagan, Dawn. P., Alfonso, V. C., Ortiz, S. O., & Dynda, A. M. (2013). Cognitive Assessment: Progress in Psychometric Theories of Intelligence, the Structure of Cognitive Ability Tests, and Interpretive Approaches to Cognitive Test Performance. In D. H. Saklofske, C. R. Reynolds, V. Schwean (Eds.), *The Oxford handbook of child psychological assessment* (pp. 239–285). Oxford University Press.

Lachman, M. E., Agrigoroaei, S., Tun, P. A., & Weaver, S. L. (2014). Monitoring Cognitive Functioning: Psychometric Properties of the Brief Test of Adult Cognition by Telephone. *Assessment, 21*(4), 404–417. https://doi.org/10.1177/1073191113508807

Lee, P., Li, P.-C., Liu, C.-H., Lin, H.-Y., Huang, C.-Y., & Hsieh, C.-L. (2021). Practice Effects, Test–Retest Reliability, and Minimal Detectable Change of the Ruff 2 and 7 Selective Attention Test in Patients with Schizophrenia. *International Journal of Environmental Research and Public Health, 18*(18), 9440. https://doi.org/10.3390/ijerph18189440

Lee, P., Lu, W.-S., Liu, C.-H., Lin, H.-Y., & Hsieh, C.-L. (2018). Test–Retest Reliability and Minimal Detectable Change of the D2 Test of Attention in Patients with Schizophrenia. *Archives of Clinical Neuropsychology, 33*(8), 1060–1068. https://doi.org/10.1093/arclin/acx123

Lee, S.-C., Chien, T.-H., Chu, C.-P., Lee, Y., & Chiu, E.-C. (2023). Practice effect and test–retest reliability of the Wechsler Memory Scale-Fourth Edition in people with dementia. *BMC Geriatrics, 23*(1), 209. https://doi.org/10.1186/s12877-023-03913-2

MacKeben, M., & Fletcher, D. C. (2011). Target Search and Identification Performance in Low Vision Patients. *Investigative Opthalmology & Visual Science, 52*(10), 7603. https://doi.org/10.1167/iovs.10-6728

Marques-Costa, C., Almiro, P. A., & Simões, M. R. (2018). Computerized Cognitive Tests (CCT) in elderly: A psychometric review. *European Review of Applied Psychology, 68*(2), 61–68. https://doi.org/10.1016/j.erap.2018.04.002

Meeker, W. Q., Escobar, L. A., & Pascual, F. G. (2021). *Statistical Methods for Reliability Data*. John Wiley & Sons.

Morgan, C., Honan, I., Allsop, A., Novak, I., & Badawi, N. (2019). Psychometric Properties of Assessments of Cognition in Infants With Cerebral Palsy or Motor Impairment: A Systematic Review. *Journal of Pediatric Psychology, 44*(2), 238–252. https://doi.org/10.1093/jpepsy/jsy068

Murray, N. P., & Janelle, C. M. (2003). Anxiety and Performance: A Visual Search Examination of the Processing Efficiency Theory. *Journal of Sport and Exercise Psychology, 25*(2), 171–187. https://doi.org/10.1123/jsep.25.2.171

Noble, S., Scheinost, D., & Constable, R. T. (2021). A guide to the measurement and interpretation of fMRI test-retest reliability. *Current Opinion in Behavioral Sciences, 40*, 27–32. https://doi.org/10.1016/j.cobeha.2020.12.012

Paiva, C. E., Barroso, E. M., Carneseca, E. C., De Pádua Souza, C., Dos Santos, F. T., Mendoza López, R. V., & Ribeiro Paiva, S. B. (2014). A critical analysis of test-retest reliability in instrument validation studies of cancer patients under palliative care: A systematic review. *BMC Medical Research Methodology, 14*(1), 8. https://doi.org/10.1186/1471-2288-14-8

Pautex, S., Berger, A., Chatelain, C., Herrmann, F., & Zulian, G. B. (2003). Symptom assessment in elderly cancer patients receiving palliative care. *Critical Reviews in Oncology/Hematology, 47*(3), 281–286. https://doi.org/10.1016/S1040-8428(03)00043-X

Portney, L. G. (2020). *Foundations of Clinical Research: Applications to Evidence-Based Practice*. F.A. Davis.

Proust-Lima, C., Amieva, H., Dartigues, J.-F., & Jacqmin-Gadda, H. (2006). Sensitivity of Four Psychometric Tests to Measure Cognitive Changes in Brain Aging-Population-based Studies. *American Journal of Epidemiology, 165*(3), 344–350. https://doi.org/10.1093/aje/kwk017

Reitan, R. M. (1958). Validity of the Trail Making Test as an Indicator of Organic Brain Damage. *Perceptual and Motor Skills, 8*(3), 271–276. https://doi.org/10.2466/pms.1958.8.3.271

Robertson, I. H., Manly, T., Andrade, J., Baddeley, B. T., & Yiend, J. (1997). *Sustained Attention to Response Task* [Dataset].APA PsycTests. https://doi.org/10.1037/t28308-000

Rollè, L., Gullotta, G., Trombetta, T., Curti, L., Gerino, E., Brustia, P., & Caldarera, A. M. (2019). Father Involvement and Cognitive Development in Early and Middle Childhood: A Systematic Review. *Frontiers in Psychology, 10*, 2405. https://doi.org/10.3389/fpsyg.2019.02405

Sánchez-Vincitore, L. V., Cubilla-Bonnetier, D., Marte-Santana, H., & Duñabeitia, J. A. (2023). Cognitive decline monitoring through a web-based application. *Frontiers in Aging Neuroscience, 15*, 1212496. https://doi.org/10.3389/fnagi.2023.1212496

Steinborn, M. B., Langner, R., Flehmig, H. C., & Huestegge, L. (2018). Methodology of performance scoring in the d2 sustained-attention test: Cumulative-reliability functions and practical guidelines. *Psychological Assessment, 30*(3), 339–357. https://doi.org/10.1037/pas0000482

Sürücü, L., & Maslakçi, A. (2020). Validity and reliability in quantitative research. *Business & Management Studies: An International Journal, 8*(3), 2694–2726. https://doi.org/10.15295/bmij.v8i3.1540

Tikhomirova, T., Malykh, A., & Malykh, S. (2020). Predicting Academic Achievement with Cognitive Abilities: Cross-Sectional Study across School Education. *Behavioral Sciences, 10*(10), 158. https://doi.org/10.3390/bs10100158

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

Van Patten, R., Iverson, G. L., Muzeau, M. A., & VanRavenhorst-Bell, H. A. (2021). Test–Retest Reliability and Reliable Change Estimates for Four Mobile Cognitive Tests Administered Virtually in Community-Dwelling Adults. *Frontiers in Psychology, 12*, 734947. https://doi.org/10.3389/fpsyg.2021.734947

VandenBos, G. R. (Ed.). (2015). *APA Dictionary of Psychology*, 2nd Edition. American Psychological Association. https://doi.org/10.1037/14646-000

Wagner, S., Helmreich, I., Dahmen, N., Lieb, K., & Tadic, A. (2011). Reliability of Three Alternate Forms of the Trail Making Tests A and B. *Archives of Clinical Neuropsychology, 26*(4), 314–321. https://doi.org/10.1093/arclin/acr024

Weir, J. P. (2005). Quantifying Test-Retest Reliability Using the Intraclass Correlation Coefficient and the SEM. *The Journal of Strength and Conditioning Research, 19*(1), 231. https://doi.org/10.1519/15184.1

Woods, D. L., Wyma, J. M., Herron, T. J., & Yund, E. W. (2015). The Effects of Aging, Malingering, and Traumatic Brain Injury on Computerized Trail-Making Test Performance. *PLOS ONE, 10*(6), e0124345. https://doi.org/10.1371/journal.pone.0124345

Zygouris, S., & Tsolaki, M. (2015). Computerized Cognitive Testing for Older Adults: A Review. *American Journal of Alzheimer's Disease & Other Dementias®, 30*(1), 13–28. https://doi.org/10.1177/1533317514522852