Sleep Medicine 149 (2027) 109312

Contents lists available at ScienceDirect

Sleep Medicine

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

Original Article

A digital cognitive battery for sleep-disordered breathing: Test–retest reliability and practice effects

Elena Richert [a,b,c,*,1], Minna Pitkänen [c,d,1], María K. Jónsdóttir [a,b,e], Erna Sif Arnardóttir [b], Samu Kainulainen [c,f], Timo Leppänen [c,d,g], Juha Töyräs [c,g,h], Henri Korkalainen [c,f,2], Kamilla Rún Jóhannsdóttir [a,b,2]

[a] Department of Psychology, Reykjavik University, Reykjavik, Iceland
[b] School of Technology, Reykjavik University Sleep Institute, Reykjavik, Iceland
[c] Department of Technical Physics, University of Eastern Finland, Kuopio, Finland
[d] Clinical Research Centre, Kuopio University Hospital, Wellbeing Services County of North Savo, Kuopio, Finland
[e] Landspitali - The National University Hospital of Iceland, Reykjavik, Iceland
[f] Diagnostic Imaging Center, Kuopio University Hospital, Kuopio, Finland
[g] School of Electrical Engineering and Computer Science, The University of Queensland, Brisbane, Australia
[h] Science Service Center, Kuopio University Hospital, Wellbeing Services County of North Savo, Kuopio, Finland

ARTICLE INFO

Keywords:
Obstructive sleep apnea
Cognitive functioning
Psychometrics
Digital cognitive assessment
Test–retest reliability
Practice effects
Floor and ceiling effects

ABSTRACT

Objective: Obstructive sleep apnea (OSA) affects cognitive performance and may contribute to long-term cognitive decline. Longitudinal assessments are needed for measuring changes in cognitive functioning in relation to OSA progression and treatment. This requires standardized, reliable cognitive assessments suitable for repeated use, but the psychometric properties of cognitive tests may vary across populations. We evaluated the suitability of a digital cognitive battery, covering various cognitive functions, for repeated administration in individuals with varying OSA severity, based on its psychometric properties.

Methods: The battery included 14 cognitive tasks. We assessed test–retest reliability, practice effects, and ceiling and floor effects between two administrations (mean interval 12 weeks) in 46 participants (mean age = 49 ± 16 years, 21 women) with a wide range of OSA severity (mean apnea-hypopnea-index (AHI) = 17.6 ± 17.7).

Results: Psychometric properties varied considerably between cognitive tasks. Test–retest reliability estimates were good or satisfactory ($ICC \geq .60$) for the Psychomotor Vigilance Test (PVT), Symbol Digit Modalities Task (SDMT), Paced Auditory Serial Addition Task (PASAT), Auditory Verbal Learning Task (AVLT), and subscores of the Tower of London Task (ToL). Practice effects were high for the PASAT in Level 1 ($p < .001, r = .63$) and the Inspection Time Task (ITT; $p < .001, r = .52$). Ceiling effects were present in the ITT and ToL, whereas intrusions in the AVLT and lapses in the PVT occurred infrequently and clustered at the lower bound.

Conclusion: Only a subset of cognitive tasks showed adequate psychometrics for repeated administration, highlighting the need for careful task selection and modification when developing cognitive assessments for repeated application in OSA populations. We include recommendations on cognitive test selection and test design for repeated assessments in OSA populations.

1. Introduction

Obstructive sleep apnea (OSA) is associated with cognitive difficulties and an increased risk for future cognitive decline [1], but the cognitive implications of OSA are still not fully understood. While most previous studies rely on cross-sectional study designs to compare individuals with OSA to healthy controls, longitudinal observations are needed to monitor the long-term impact of disorder progression and therapeutic interventions on cognitive health [2,3].

Longitudinal studies require cognitive measures that can be standardized and produce reliable and interpretable results across repeated assessments. However, the current field of sleep research lacks

* Correspondence to: Department of Psychology, Reykjavik University, Menntavegur 1, 102 Reykjavík, Iceland.
E-mail addresses: elenar@ru.is, erichert@uef.fi (E. Richert).
[1] These authors share first authorship.
[2] These authors share last authorship.

https://doi.org/10.1016/j.sleep.2026.109312
Received 18 March 2026; Received in revised form 5 October 2026; Accepted 8 October 2026
Available online 16 October 2026
1389-9457/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

consensus on how cognitive functions should be measured consistently across studies [4], particularly in longitudinal assessments. Previous research on cognitive functioning in OSA has primarily focused on vigilance [5–9], verbal memory [10–15], and executive functions [10,15–19], reporting a relatively consistent link between OSA and poorer performance. In contrast, other cognitive functions, including visuoconstructional abilities, nonverbal memory, social cognition, and basic perceptual functions, remain underexplored [4]. Moreover, different studies might use the same cognitive test to assess distinct cognitive functions [4]. Together with differences in test administration, these factors complicate comparisons across studies, leading to an incomplete picture of the cognitive effects of OSA.

In addition, cognitive tests used to assess differences between individuals with OSA and healthy controls in cross-sectional studies may not necessarily be suitable for repeated assessments in longitudinal designs. A test that is sensitive to differences between individuals may not be sufficiently reliable or sensitive to measure true changes within the same individual over time. Therefore, cognitive tests intended for repeated assessments in longitudinal studies require a specific evaluation of their psychometric properties. Furthermore, psychometric properties are sample-dependent and may vary across populations [20]. Consequently, findings from healthy populations cannot be automatically generalized to individuals with OSA, and a sample-specific evaluation is required.

One of the most important psychometric properties for longitudinal designs is test–retest reliability, which indicates a measurement tool's consistency and reproducibility of results across repeated assessments under identical conditions [2]. An important factor affecting test–retest reliability is practice effects, which characterize a change in test performance due to test familiarity rather than true changes in cognitive performance [21–23]. Practice effects can arise from memory of individual test items or test-taking strategies and are influenced by various factors, such as participant characteristics, retest intervals, and the choice of cognitive tests [22–24]. If practice effects are not considered in the cognitive assessment, they can conceal cognitive decline or lead to an overestimation of treatment effects [25–27]. Importantly, cognitive tests with systematic practice effects, which affect all test-takers similarly, can still show high test–retest reliability, whereas unsystematic practice effects will lower test–retest reliability [24]. Additionally, floor and ceiling effects must be taken into account. Floor effects occur when many individuals achieve the lowest or nearly lowest scores [28,29], for example, when a test is too difficult or the outcome measure is rarely observed. Vice versa, ceiling effects occur when many individuals achieve the highest or nearly highest scores, for example, when a test is too easy, or an outcome measure is highly prevalent [30,31]. This compresses the score distribution at the lower or upper bound, decreasing the true variability between individuals. Consequently, this leads to an inaccurate depiction of an individual's full capabilities, may conceal practice effects [28,32], and result in lower test-retest reliability [33]. It is therefore necessary to select cognitive tests with high test–retest reliability, minimal practice effects, and an appropriate level of difficulty.

One form of administering cognitive assessments that has been repeatedly proposed for regular testing [34–37], and has gained popularity in recent years [37], is the use of digital versions of traditional paper-based tests. Cognitive assessments conducted through digital means allow precise stimulus presentation and can reduce human error and examiner-dependent variability through automated scoring, data collection, and documentation. Moreover, the exact measurement of reaction times and solution times provides valuable metrics for understanding cognitive processes. Thus, digital tests can be highly standardized for test presentation and scoring and improve accessibility to cognitive assessments. However, digitalization of cognitive assessments does not guarantee suitability for longitudinal studies, and general considerations and limitations of cognitive testing remain, including test–retest reliability, practice effects, and ceiling and floor effects [38,39].

We have previously reported associations between cognitive performance and sleep and OSA-related parameters in this dataset in Ferretti et al. [40] and Thorisdottir et al. [41]. In this study, we propose and evaluate a digital cognitive test battery for individuals with OSA. The cognitive test battery includes assessments of multiple cognitive domains, including attention, perceptual-motor skills, memory and learning, executive functions, and social cognition. We investigate the potential of cognitive tasks for repeated administration in longitudinal assessments, considering test–retest reliability, practice effects, as well as floor and ceiling effects. We expect differences in the psychometric properties among the cognitive tasks and provide recommendations for adjusting the battery accordingly. Altogether, we aim to contribute to the development of a comprehensive digital tool for repeated assessments of cognitive functioning in individuals with OSA.

2. Methods

2.1. Research design and procedure

This study got ethical approval from the National Bioethics Committee and Data Protection Authority of Iceland (VSN number 21-070). Data were collected as part of the Sleep Revolution project [42] between November 2021 and May 2022 at Reykjavik University in Reykjavik, Iceland, and managed according to Sveinbjarnarson et al. [43] to ensure data standardization and quality control. All participants signed an informed consent letter before participation. Participants were not financially reimbursed, but received a sleep report at the end of the study. The sleep report included a summary of the participant's sleep assessment results, including sleep architecture, body position, respiratory and cardiac parameters, oxygen saturation, and results from the Epworth Sleepiness Scale (ESS), accompanied by educational information explaining the sleep-related parameters and general sleep hygiene, as well as guidance on whether consultation with their general practitioner was recommended based on the results.

The study was advertised locally and through the project's participation registry to those with suspected or confirmed OSA and healthy participants. Online questionnaires were sent out via the Research Electronic Data Capture (REDCap) platform [44] to collect information on demographics, comorbidities, and sleep complaints, including the ESS to assess daytime sleepiness [45]. Participants with unstable terminal illness, and hearing or visual impairment that could not be corrected to normal, were not considered for study participation. Another exclusion criterion was ongoing OSA treatment (e.g., positive airway pressure (PAP) treatment), as such treatment can improve cognitive performance [46], thereby complicating the distinction between treatment-related changes and the psychometric properties of the selected cognitive tests. This resulted in a total sample of 46 participants (21 women, 25 men) with an average age of 49 years (SD = 16, range 19–79). Fourteen participants held a postgraduate degree (MS/MA/MBA, PhD), 13 held an undergraduate degree (BS/BA), six held a vocational/technical degree, four had graduated from upper secondary school (Icelandic "menntaskóli"), four had graduated from compulsory lower secondary school (Icelandic "grunnskóli"), and five did not disclose their education level.

Participants attended two visits for baseline and retest. The baseline consisted of a cognitive assessment under standardized conditions at Reykjavik University and a subsequent three-night self-applied polysomnography (PSG; Nox Medical, A1s, Reykjavík, Iceland) at home. The retest visit took place at Reykjavik University about 12 weeks later (mean interval 93 days, SD = 16, range 76–154) to repeat the cognitive assessment. Both cognitive assessments were scheduled for 10 am or 12 pm to control for time-of-day effects [47]. No intervention was administered throughout the 12 weeks, and participants were encouraged to maintain their lifestyles as stable as possible. One participant left the second visit early due to personal time restrictions and therefore completed only the first nine tests of the cognitive test battery.

Table 1
Cognitive test battery and analyzed variables.

Task	Main domain/processes	Analyzed variable(s)
Simple Reaction Time Task (SRTT)	Attention, motor speed	Mean reaction time (ms)
Inspection Time Task (ITT)	Processing speed, perceptual speed, selective attention	Average correct responses (proportion)
Auditory Verbal Learning Task (AVLT)	Memory, verbal memory	Immediate recall score Delayed recall score
Psychomotor Vigilance Task (PVT)	Attention, vigilance, sustained attention	Mean reaction time (ms) Mean reaction time <500 ms Number of lapses
Symbol Digit Modalities Task (SDMT)	Processing speed, psychomotor speed, visual scanning	Number of correct entries
Tower of London (ToL)	Executive functions, planning, problem-solving	Mean solution time (ms) Mean first-move latency (ms) Total score
Corsi Block Tapping Task (CBTT) – Forward Version	Executive functions, visuospatial working memory	Block span
Corsi Block Tapping Task (CBTT) – Backward Version	Executive functions, visuospatial working memory	Block span
Visual Design Learning Task (VDLT)	Memory, non-verbal memory, visuoconstruction	Hit rate (proportion)
Iowa Gambling Task (IGT)	Executive functions, reasoning, decision-making	Amount of coins won
Penn Emotional Recognition Task (ER-40)	Social cognition, emotional recognition	Correct answers (proportion)
Judgment of Line Angle and Position (JLAP-15)	Perception, visuoperceptual ability	Total correct answers
Task Switching Paradigm (TSP)	Executive functions, cognitive flexibility	Latency switch cost for crosstalk trials (ms) Latency switch cost for non-crosstalk trials (ms) Accuracy switch cost for crosstalk trials (proportion) Accuracy switch cost for non-crosstalk trials (proportion)
Paced Auditory Serial Addition Task (PASAT)	Executive functions, fast-paced working memory	Number of correct responses in level 1 Number of correct responses in level 2

ms = milliseconds.

2.2. Digital cognitive test battery

Up to two participants simultaneously completed the digital cognitive test battery in a designated research room, with a researcher present at all times. Both participant setups were identical regarding the computer, computer screen, mouse, keyboard, noise-canceling headphones with standardized volume levels, and desk privacy screens. The digital cognitive test battery was presented in Icelandic and consisted of 14 tasks, based on the recommendations of Jóhannsdóttir et al. [4], and took on average 96 min (SD = 17, range 67–159) to complete. The cognitive tasks were administered in Inquisit (Version 6.5.1) [48], mainly based on the Millisecond test library, and modified as needed. All tasks were presented with a light gray background, except both versions of the Corsi Block Tapping Task, which had a black background. The descriptions in Table 1 follow the order of task presentation, and detailed information on task design is provided in the Supplementary Material A.

2.3. Restful levels

Participants rated their level of restfulness before and after completing the cognitive test battery by responding to the statement "I feel as rested as usual" on a 5-point Likert scale ranging from "Strongly disagree" to "Strongly agree". We analyzed changes before and after completion of the cognitive assessment, as well as differences between baseline and retest.

2.4. Sleep study

Participants underwent a three-night self-applied PSG at home to assess OSA severity. The setup has previously been described by Ferretti et al. [49] and validated by Rusanen et al. [50] and Kainulainen et al. [51]. Expert sleep technologists scored the PSGs according to the American Academy of Sleep Medicine (AASM) manual, version 2.6 [52]. Nights with a total sleep time below 4 hours were considered invalid and not included in the analysis. The apnea-hypopnea index (AHI), total sleep time (TST), and sleep efficiency (SE) were calculated as an average over all available valid nights for each participant. OSA severity was based on the traditional AHI thresholds (no OSA: AHI < 5, mild: 5 ≤ AHI < 15, moderate: 15 ≤ AHI < 30, severe: AHI ≥ 30 events/hour) [53]. Two participants had no valid nights, resulting in missing sleep data.

2.5. Statistical analysis

We defined the outliers in the cognitive task scores as values more than 1.5 times the interquartile range away from the 25th and 75th percentiles and excluded them from the analysis, except for the CBTT. Only zero values were considered as outliers for the CBTT forward and backwards due to the limited block span range.

We conducted Wilcoxon signed-rank tests in the PSPP software (GNU PSPP, Version 1.6.2-g78a33a, Boston, MA) to analyze differences in test scores between the two cognitive assessments, where positive values indicate a higher baseline than retest score. The effect size (r) was calculated as the absolute value of the z statistic divided by the square root of the number of participants. We analyzed the consistency between the two assessments for each subscore using the intraclass correlation coefficient (*ICC*, two-way mixed-effects model, single measurement, absolute agreement) in R 4.3.1 [54] and the *irr* package [55]. We assessed the effects of the test–retest interval and age on practice effects using a multiple linear regression in R Statistics with the *lm* function. Due to scale differences across the cognitive task variables, we used a standardized change score (the difference between the retest and baseline scores, divided by the standard deviation of the differences) as the dependent variable. The independent variables were the time interval between baseline and retest, and age. We report the p-value for the overall model. We conducted a linear mixed model with the *lmerTest* package [56] in R Statistics to analyze changes in restfulness before and after completion of the neurocognitive test battery (timepoint), and differences between baseline and retest (session). The dependent variable was the restfulness level, the random effect was the participant, and fixed effects were timepoint, session, and their interaction. As the interaction was not significant, it was removed in the final model to increase the degrees of freedom for estimating the main effects. The significance level was set to α <.05. Corrections were used for multiple comparisons for the Wilcoxon signed-rank tests and

Table 2
Comparison of results at baseline and retest.

Task parameter	ICC (95% CI)	Mean 1 (SD)	Mean 2 (SD)	Wilcoxon test *p*-value	Effect size	*N*
SRTT, mean reaction time	0.36 (0.04 – 0.62)	292 (48)	293 (35)	0.718	0.06	36
ITT, proportion of correct	0.38 (0.07 – 0.62)	0.88 (0.14)	0.94 (0.05)	**0.001***	0.52	37
AVLT, recall score	0.74 (0.55 – 0.86)	45.02 (10.76)	48.36 (12.59)	0.005	0.43	44
AVLT, delayed recall score	0.69 (0.46 – 0.82)	10.30 (2.91)	11.27 (2.86)	0.007	0.41	44
AVLT, mean intrusions	0.44 (0.18 – 0.65)	0.44 (0.49)	0.58 (0.57)	0.099	0.25	43
PVT, mean reaction time	0.84 (0.68 – 0.91)	325 (47)	337 (52)	0.034	0.33	41
PVT, mean reaction time ≤ 500ms	0.92 (0.85 – 0.98)	326 (47)	333 (45)	0.015	0.36	45
PVT, lapses	0.46 (0.15 – 0.68)	1.61 (2.02)	1.83 (2.60)	0.847	0.03	36
SDMT, number of correct entries	0.87 (0.77 – 0.93)	57.17 (24.28)	58.14 (24.79)	0.495	0.10	42
ToL, mean solution time	0.76 (0.60 – 0.86)	16.49 (4.58)	15.56 (4.50)	0.049	0.29	45
ToL, mean first-move latency	0.64 (0.44 – 0.79)	9.54 (3.11)	8.91 (3.52)	0.268	0.17	44
ToL, total score	0.09 (−0.17 – 0.36)	32.38 (3.83)	34.33 (1.87)	0.005	0.45	39
CBTT, forward, block span	0.17 (−0.10 – 0.43)	5.62 (1.25)	6.38 (1.14)	0.004	0.46	39
CBTT, backward, block span	0.27 (−0.04 – 0.54)	6.23 (0.86)	6.23 (1.07)	0.923	0.02	40
VDLT, hit rate	0.54 (0.25 – 0.74)	0.86 (0.11)	0.90 (0.08)	0.005	0.44	41
IGT, amount of coins won	−0.00 (−0.27 – 0.28)	1444 (1101)	1871 (1080)	0.021	0.34	45
ER-40, proportion of correct	0.54 (0.28 – 0.72)	0.83 (0.06)	0.83 (0.09)	0.783	0.04	42
JLAP-15, total correct	0.27 (−0.03 – 0.52)	13.31 (3.27)	13.38 (3.32)	0.580	0.08	45
TSP, ACC, switch cost C	0.28 (−0.02 – 0.54)	−0.04 (0.06)	−0.05 (0.07)	0.416	0.13	41
TSP, RT, switch cost C	0.61 (0.36 – 0.77)	0.34 (0.22)	0.35 (0.18)	0.617	0.08	38
TSP, ACC, switch cost N	0.31 (−0.02 – 0.57)	−0.04 (0.06)	−0.07 (0.07)	0.047	0.32	38
TSP, RT, switch cost N	0.62 (0.40 – 0.77)	0.28 (0.23)	0.29 (0.19)	0.852	0.03	44
PASAT1, number of correct	0.66 (0.30 – 0.83)	31.43 (14.47)	38.77 (15.44)	<**0.001***	0.63	44
PASAT2, number of correct	0.81 (0.66 – 0.89)	36.88 (18.14)	40.71 (21.51)	0.042	0.31	42

ACC = Accuracy, AVLT = Auditory verbal learning task, CBTT = Corsi block tapping task, CI = Confidence interval, ER-40 = Penn emotional recognition task, ICC = Intraclass correlation coefficient, IGT = Iowa gambling task, ITT = Inspection time task, IQR = Interquartile range, JLAP-15 = Judgment of line angle and position, *N* = Number of participants after removing outliers, PASAT1 = Paced auditory serial addition task level 1, PASAT2 = Paced auditory serial addition task level 2, PVT = Psychomotor vigilance task, RT = Reaction time, SD = standard deviation, SDMT = Symbol digit modalities task, SRTT = Simple reaction time task, ToL = Tower of London, TSP = Task switching paradigm, VDLT = Visual design learning task.

* Denotes statistical significance ($p - value < .002$).

multiple linear regressions. Following the Bonferroni approach, α was divided by the number of analyzed task variables (n = 24), resulting in an adjusted significance level of α =.002.

3. Results

3.1. Test–retest reliability

Test–retest reliability varied substantially, with wide 95% confidence intervals for the majority of cognitive tasks. ICC estimates ranged from 0 for the IGT (amount of coins won) to .92 for the PVT (mean reaction time < 500 ms, Table 2). Based on the ICC estimates, good reliability (≥.70) was achieved by PASAT (level 2, number of correct responses), AVLT (recall score), PVT (mean reaction time, mean reaction time < 500 ms), SDMT (count of correct responses), and ToL (mean solution time). ICC estimates were satisfactory (.60–.69) for TSP (RT switch cost C, RT switch cost N), ToL (mean first-move latency), PASAT (level 1, number of correct responses), and AVLT (delayed recall score). Cognitive tasks with low ICC estimates (<.60) include CBTT backward version (block span) and CBTT forward version (block span), VDLT (hit rate), AVLT (mean intrusions), PVT (number of lapses), SRTT (mean reaction time), JLAP-15 (total correct), ITT (average proportion of correct responses), ToL (total score), ER-40 (proportion of correct responses), IGT (amount of coins won), and TSP (accuracy switch cost C, accuracy switch cost N).

3.2. Practice effects

Several cognitive tasks differed significantly ($p < .05$) between the two administrations (Table 2). The participants improved significantly in the CBTT forward version (0.76 higher block span), ITT (6% more proportion of correct responses), IGT (427 more coins in amount of coins won), PASAT (7.3 more correct responses in level 1 and 3.8 more correct responses in level 2), VDLT (4% higher hit rate), ToL (0.93 s faster mean solution time and 2.0 higher total score), and AVLT (3.3 more items in recall scores). In contrast, performance worsened for the PVT (mean reaction time and mean reaction time < 500 ms) with a difference of 12 ms and 7 ms, and the TSP (accuracy, switch cost N) with a difference of 3%. After Bonferroni correction, resulting in an adjusted significance level of .002, the differences remained significant for the ITT (average proportion of correct responses, $r = .52$) and PASAT level 1 (number of correct responses, $r = .63$) with large effect sizes. Other cognitive task parameters showed no significant difference between sessions.

Multiple linear regressions examining the effects of test–retest interval and age on practice effects were not significant for any of the cognitive tasks with *p*-values ranging from .185 to .994. Detailed results can be found in Supplementary Material B.

3.3. Floor and ceiling effects

Visual inspection of the scatterplots (Fig. 1) suggests potential floor and ceiling effects for several cognitive tasks. A clustering at the lower bound is prominent for the AVLT (mean intrusions), and the PVT (number of lapses). Conversely, a clustering at the upper bound is visible for the ITT (proportion of correct answers) and the ToL (total score).

3.4. Restful levels

Data on restful levels were available from 26 participants from both timepoints (before and after the cognitive assessment) from both sessions (baseline and retest), and from 20 participants for both timepoints from either one session. Restful levels were on average 3.1 points ($SD = 1.3$) before and 2.7 points ($SD = 1.1$) after completion of the cognitive test battery at baseline, and 3.3 points ($SD = 1.4$) and 3.2 points ($SD = 1.2$) at retest before and after completing the cognitive test battery, respectively. While the average restfulness level tended to be lower after the cognitive assessment, there was no significant main effect of timepoint ($p = .053$) or session ($p = .223$). Instead, restful levels varied significantly between participants ($SD = 1, p < .001$).

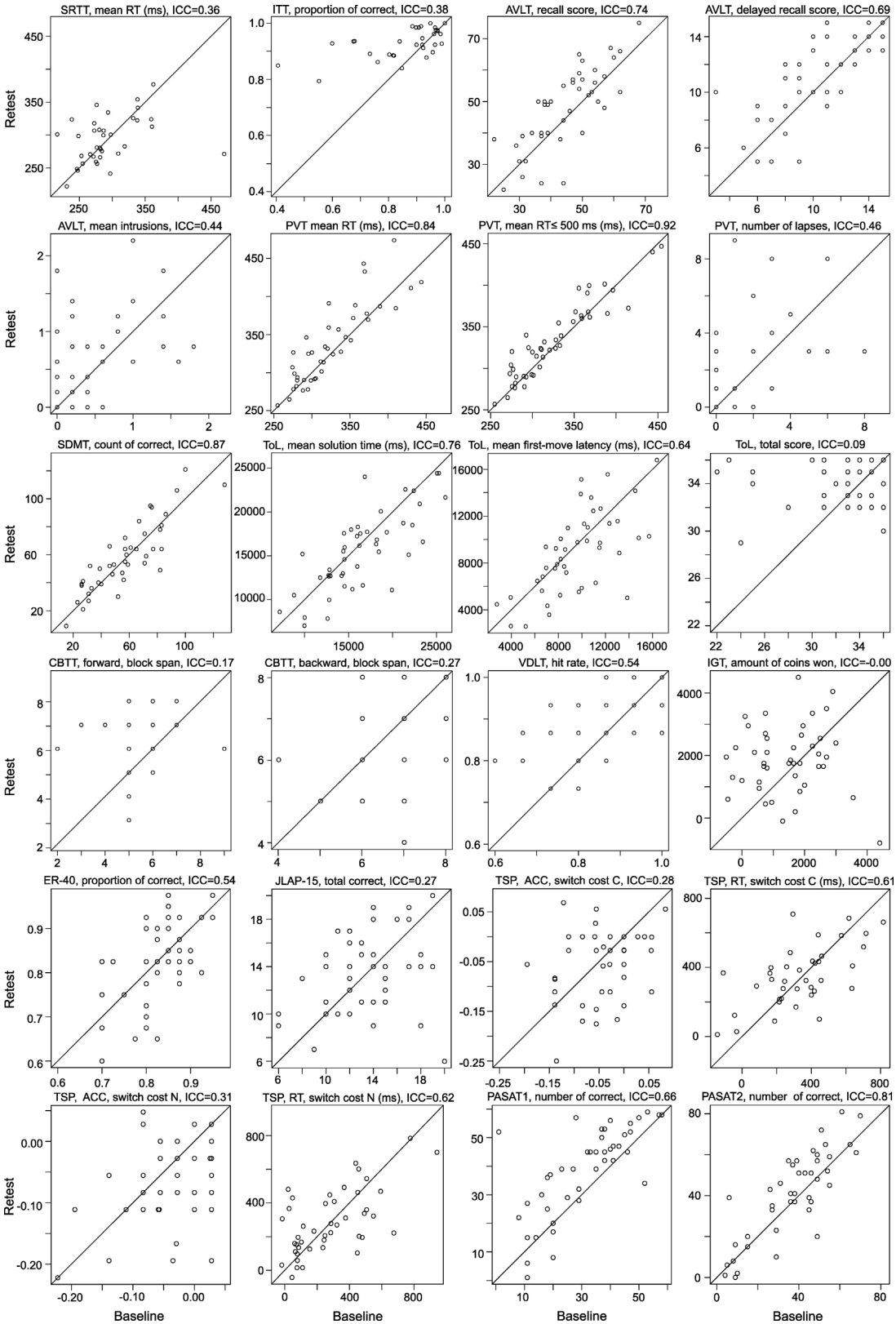


Fig. 1. Scatter plots including the lines for the absolute agreement and the intraclass correlation coefficients (ICC). Outliers are excluded. ACC = accuracy, AVLT = Auditory verbal learning task, CBTT = Corsi block tapping task, ER-40 = Penn emotional recognition task, ICC = Intraclass correlation coefficient, IGT = Iowa gambling task, ITT = Inspection time task, JLAP-15 = Judgment of line angle and position, PASAT1 = Paced auditory serial addition task level 1, PASAT2 = Paced auditory serial addition task level 2, PVT = Psychomotor vigilance task, RT = Reaction time, SDMT = Symbol digit modalities task, SRTT = Simple reaction time task, ToL = Tower of London, TSP = Task switching paradigm, VDLT = Visual design learning task.

3.5. Sleep characteristics

The participants showed a wide range of OSA severity. Based on the AHI (M = 17.6, SD = 17.7, range 0.4–74.7), 12 participants were classified as no OSA, 12 as mild OSA, 14 as moderate OSA, and six as severe OSA. Total sleep time was on average 6.4 h (SD = 0.8, range 4.5–8.3) and mean sleep efficiency was 82.5% (SD = 8.1, range 63.1–95.1).

ESS scores ranged from 0 to 20 points, with an average of 8.7 (SD = 4.6). Based on a cut-off of 10 points, 31 participants reached scores at or below the threshold for excessive daytime sleepiness, and 15 participants scored above the cutoff.

4. Discussion

In this study, we evaluated the test–retest reliability, practice effects, as well as floor and ceiling effects of a novel digital cognitive battery for OSA covering multiple cognitive domains, including attention, perceptual-motor skills, memory and learning, executive functions, and social cognition. We aimed to identify cognitive tasks suitable for longitudinal assessments with repeated measurements for individuals with OSA. Based on our findings, we propose recommendations for selecting and adapting cognitive tests.

The SDMT was the only cognitive task with good to high test–retest reliability, no practice effects, and no floor or ceiling effects, as also demonstrated in previous studies [57,58]. Other cognitive tasks with overall satisfactory to good psychometrics include the PVT, PASAT, and AVLT, but require careful consideration. For the PVT, test–retest reliability was satisfactory to high for both mean reaction time measures, particularly mean reaction times < 500 ms, and low for the number of lapses. Furthermore, mean reaction times differed with a medium effect size between the two assessments when not corrected for multiple comparisons, being slower at retest. However, the difference can be considered negligible, and previous research has demonstrated consistent PVT performance over time [59,60]. In contrast, the test–retest reliability for PVT lapses spanned from low to satisfactory. This is likely due to the overall low number of lapses in our study, whereas other studies often report frequent lapses in participants with OSA [5,6], as well as good test–retest reliability for PVT lapses [61]. Notably, our sample differed from many OSA studies in its wide range of age and OSA severity, including many individuals with low or no OSA, as our study participants were recruited from the general population. In contrast, many clinical studies focus almost exclusively on older participants with moderate or severe OSA [62]. Overall, PVT reaction times, particularly < 500 ms, might be a more reliable outcome measure than lapses in younger and less severe OSA cases.

For the PASAT, test–retest reliability ranged from low to good across interstimulus intervals (ISI), favoring the shorter ISI condition, and moderate to large practice effects were present. Practice effects are well-known for the PASAT [63–65], but tend to plateau after repeated testing [63,65]. Therefore, it is recommended to administer the PASAT several times before determining a baseline, but the assessment time required to establish a baseline varies between patient groups [65]. Furthermore, practice effects differ between ISIs [66], as was also present in our study. Practice effects were larger for the longer ISI condition and remained significant after adjusting for multiple comparisons. Currently, neither the optimal number of assessments to identify a baseline nor the best ISI for participants with OSA has been determined.

The AVLT demonstrated low to good test–retest reliability for immediate and delayed recall, with higher reliability for the immediate recall, and medium-sized practice effects that were no longer significant after correction for multiple comparisons. Practice effects are often reported for verbal learning tests [22,63], but can be reduced with alternate forms [22], although currently not widely available [67]. The low to satisfactory test–retest reliability for the AVLT intrusions was most likely caused by their minimal occurrence in our sample. Although verbal learning tests are regularly used in studies on OSA [4], intrusions are usually not reported. While intrusions are minimal in healthy participants [68], they have high predictive value in differentiating between normal ageing and cognitive decline [69], and can be a useful parameter for longitudinal assessments of cognitive functioning in OSA.

One task showing promising results, but also ceiling effects, was the ToL. The mean solution time for correctly solved items showed satisfactory to high test–retest reliability, and although small practice effects were present when not adjusted for multiple comparisons, the approximately 1-sec difference might be negligible. Notably, repeated use of the same ToL version may result in larger practice effects across multiple assessments. The mean first-move latency showed low to good test–retest reliability and no practice effects. However, participants were not explicitly instructed to plan their moves before attempting to solve the task, which may have influenced the first-move latency. Therefore, the influence of different types of instructions on psychometric properties should be assessed in future research. Lastly, the total score showed one of the lowest test–retest reliabilities in this study, likely due to pronounced ceiling effects. Overall, these are a well-recognized limitation of the classical ToL [70], and have led to different adjustments to increase test difficulty, including time restrictions, items requiring more moves, or a combination of both [71–73], and the use of parallel-test versions [74]. However, there is currently no recommendation on which design of the ToL is most appropriate for participants with OSA.

The remaining cognitive tasks, including TSP, VDLT, ER-40, ITT, SRTT, JLAP-15, both CBTT forward and backwards, and IGT, produced insufficient psychometrics. Test–retest reliability spanned a wide range and estimates were low. Additionally, moderate to large practice effects emerged in the TSP, VDLT, ITT, CBTT forward span, and IGT, although they remained significant only for the ITT after adjusting for multiple comparisons. Additionally, the ITT showed ceiling effects. These findings are mostly consistent with previous research. Low test–retest reliability has been reported for the VDLT in children [75], ER-40 [76,77], ITT [78], CBTT [79], and IGT [80,81]. Practice effects have been found for the VDLT in children [75], the ITT [78,82,83], CBTT [79,84], and IGT [80,81,85]. However, practice effects in the IGT are expected for control groups and are often used to distinguish various clinical groups from controls [85]. In addition, ceiling effects have previously been found for the ITT [78]. Notably, alternative test versions or scoring methods of some of these cognitive tests have been shown to improve test–retest reliability and reduce practice effects, such as a mobile version of the ITT [86], a JLAP version with 13 lines tested in young college students [87], the CBTT when using the mean span across multiple trials instead of the maximum block span [88], or the IGT when using a statistical modeling approach to increase its test–retest reliability proposed by Sullivan-Toole et al. [81]. Nevertheless, many of the cognitive tasks included in this study lack psychometric literature on repeated testing or are difficult to compare because of methodological differences. This is especially prominent for the SRTT and TSP due to their highly heterogeneous methodology in the number of practice and test trials, ISIs, stimulus types, and response types [89]. Furthermore, the TSP is highly sensitive to measurement noise [90,91], and it is advisable to increase the number of trials to obtain more stable results. Additionally, alternative scoring methods that combine reaction time and accuracy into a single parameter, as proposed by Hughes et al. [90] and Draheim et al. [92], have been shown not only to be more reliable than the traditional TSP measures reported in our study, but also to increase TSP validity.

4.1. Summary

Overall, these findings show that not all tasks within the tested cognitive battery are equally suitable for repeated administration in individuals with OSA. The cognitive tasks with the most robust psychometric properties included the SDMT, PVT, PASAT, and AVLT,

Table 3
Psychometric findings and recommendations for repeated administration of cognitive tasks in OSA populations.

Task	Main domain/processes	Main psychometric findings	Recommendations for repeated administration in OSA	Recommendations for potential adaptations and future research
SRTT	Attention, motor speed	Low test-retest reliability No practice effects No floor or ceiling effects	Not recommended in its current form	Requires further research on the optimal number of test trials, ISI, stimulus, and response types
ITT	Processing speed, perceptual speed, selective attention	Low test-retest reliability Large practice effects Shows ceiling effects	Not recommended in its current form	Investigate alternate versions, including adjustments in stimuli, response types, and scoring methods
AVLT	Memory, verbal memory	Satisfactory – good test-retest reliability Medium practice effects Few intrusions	Immediate and delayed recalls are strong candidates, intrusions might potentially be disregarded in younger and less severe OSA cases	Use alternate forms to reduce practice effects, develop more alternate forms, evaluate psychometric properties of intrusions in older and more severe OSA samples
PVT	Attention, vigilance, sustained attention	High test-retest reliability Negligible practice effects Low occurrence of lapses in this sample	Strong candidate, RT might be more reliable than lapses in younger and less severe OSA cases	Explore whether feedback on RT might influence psychometric properties
SDMT	Processing speed, psychomotor speed, visual scanning	Excellent test-retest reliability No practice effects No floor or ceiling effects	Strong candidate	Explore whether parallel versions might be required for multiple (n > 2) assessments
ToL	Executive functions, planning, problem-solving	Good test-retest reliability of solution time, satisfactory for mean first-move latency, and low for total score Moderate practice effects Shows ceiling effects for total score	Potential candidate, mean solution time and mean first-latency might be more suitable measures than total score	Use parallel-test versions, increase test difficulty
CBTT	Executive functions, visuospatial working memory	Low test-retest reliability Moderate practice effects No floor or ceiling effects	Not recommended in their current form	Explore other scoring methods, such as mean block span across multiple trials
VDLT	Memory, non-verbal memory, visuoconstruction	Low test-retest reliability Moderate practice effects No floor or ceiling effects	Not recommended in its current form	Use traditional analogue tests, evaluate alternate administration forms, such as different input devices, and develop algorithms for automated image recognition to simplify scoring procedures
IGT	Executive functions, reasoning, decision-making	Low test-retest reliability Small practice effects No floor or ceiling effects	Not recommended in its current form	Explore different scoring methods, e.g. through statistical modeling approaches, investigate whether practice effects can be used to identify clinical groups
ER-40	Social cognition, emotional recognition	Low test-retest reliability No practice effects No floor or ceiling effects	Not recommended in its current form	Investigate psychometrics of other metrics, such as recognition success per emotion type or intensity
JLAP-15	Perception, visuoperceptual ability	Low test-retest reliability No practice effects No floor or ceiling effects	Not recommended in its current form	Explore alternate versions, such as using 13 lines, possibly increase the number of trials
TSP	Executive functions, cognitive flexibility	Satisfactory test–retest reliability for switch-cost RT, low for switch-cost ACC Moderate practice effects for one switch-cost ACC condition No floor or ceiling effects	Switch-cost RT may be a potential candidate, switch-cost ACC is less reliable	Increase the number of trials, consider alternate scoring methods, further validation of task parameters is needed
PASAT	Executive functions, fast-paced working memory	Satisfactory – good test-retest reliability Practice effects depend on ISI Floor effects for very short ISI	May be suitable, but practice effects and ISI should be considered	Administer multiple times before defining baseline, identify suitable ISI and minimum repetitions for baseline for populations with OSA

ACC = Accuracy, AVLT = Auditory verbal learning task, CBTT = Corsi block tapping task, ER-40 = Penn emotional recognition task, IGT = Iowa gambling task, ITT = Inspection time task, JLAP-15 = Judgment of line angle and position, PASAT = Paced auditory serial addition task, PVT = Psychomotor vigilance task, RT = Reaction time, SDMT = Symbol digit modalities task, SRTT = Simple reaction time task, ToL = Tower of London, TSP = Task switching paradigm, VDLT=Visual design learning task.

measuring processing speed, vigilance, working memory, and verbal memory, respectively. The lowest test–retest reliability was observed among cognitive tasks that measure executive functions, nonverbal memory, and perceptual skills. Practice effects were most prominent among cognitive tasks assessing executive functions and processing speed. The main results of each cognitive task, recommendations for their use in an OSA population, potential adaptations, and requirements for further research are summarized in Table 3. Importantly, these recommendations concern the suitability of the cognitive tasks for repeated cognitive assessments to monitor possible disorder progression or treatment effects in longitudinal settings, but not whether these tasks are suitable for assessing differences between individuals with OSA and healthy controls in cross-sectional designs.

4.2. Methodological considerations and limitations and future research

To the best of our knowledge, this is the first study to assess test–retest reliability, practice effects, and floor and ceiling effects across various cognitive functions and in a demographically and clinically diverse population with and without OSA. This sample is more likely to represent individuals who are referred for diagnostic sleep evaluations

and general population assessment of OSA, supporting the generalizability of our findings, as most OSA research is focused on older men with more severe disease [62,93]. We also increased the ecological validity of sleep assessments by conducting PSG studies over three consecutive nights in the participant's home, reducing the first-night effect and bias from night-to-night variability [94–96]. Furthermore, we minimized circadian rhythm effects on cognitive performance through consistent timing of the cognitive assessment [47], and the digital administration of the cognitive battery allowed for a highly standardized assessment.

However, several limitations of this study must be acknowledged. Due to the small sample size, we did not perform subgroup analyses. Although we analyzed the impact of age on practice effects, larger samples of at least 300–400 participants are recommended for test–retest reliability assessments [97], and psychometric properties should be evaluated across age groups, education levels, OSA severity levels, and potentially other common OSA symptoms, such as daytime sleepiness [98,99]. For example, more severe OSA may be associated with greater performance variability and reduced learning, potentially affecting test–retest reliability and practice effects. Similarly, floor and ceiling effects might differ between groups: More severe OSA might lead to increased floor effects on tests with a high cognitive load demand, such as the PASAT, but lower floor effects in the PVT lapses and possibly AVLT intrusions, whereas ceiling effects may be less pronounced in other tests, such as the ITT or ToL. Performing such subgroup analyses may therefore strengthen the clinical utility of cognitive tests and determine differences in psychometric properties and the suitability of individual cognitive tests for specific subgroups. Moreover, this may help to identify individuals at increased risk of more severe cognitive dysfunction [100], and inform treatment decisions, similar to other clinical groups where low practice effects have been associated with future cognitive decline [101,102], neurodegeneration biomarkers [27], and treatment response [23].

Another critical aspect is that we did not control for medication or intake of stimulants, such as caffeine or nicotine, which might have influenced task performance. Additionally, participants underwent a sleep study only after the cognitive assessment at baseline. Therefore, it cannot be determined how the previous night's sleep influenced task performance, nor whether sleep parameters remained consistent between the two assessments. Nevertheless, the three-night PSG likely presents a more reliable baseline compared to the traditional one-night assessment, and it is unlikely that OSA-related parameters, such as the AHI, changed during the approximately 12-week test–retest interval with no intervention. Nevertheless, future studies should ideally conduct a sleep study before each cognitive assessment, ideally over multiple nights, and within a standardized test–retest interval.

Several limitations also arise from the cognitive test battery. While using digital test versions over traditional analogue ones has multiple advantages, primarily high standardization and the measurement of additional parameters, there is an ongoing debate about their direct comparability with traditional paper-based versions in terms of clinical validity and the current lack of psychometric data [35,103,104]. Future studies should therefore evaluate the comparability of digital and traditional pen-and-paper-based cognitive tests, provide information on psychometric properties and norms of digital test versions, and research the validity of digital tests. Additionally, graphomotor control, such as drawing, differs between pen-and-paper and digital formats and may even vary across types of digital input devices [105,106]. Moreover, digital drawings may complicate scoring due to irregularities in graphomotor performance, as observed in the VDLT in our study. This could be facilitated and automated in the future through image recognition algorithms. Until then, researchers and clinicians may consider including the traditional pen-and-paper-based drawing tasks. Furthermore, familiarity and acceptance of digital devices can affect cognitive performance on digital tests, particularly among older participants [37], and should be assessed in future studies.

Another aspect was that the cognitive tasks were presented in the same order, which could introduce test-order effects [107]. The order of test administration within a cognitive battery can influence overall performance and individual test performance, and thereby contribute to additional test errors [108]. This is especially the case if the same cognitive processes are measured consecutively. However, cognitive domains are not independent of one another, more complex cognitive functions depend on more basic cognitive processes, and individual cognitive tests do not measure a single, isolated cognitive process [109]. Therefore, potential test-order effects should be carefully considered when designing and administering a cognitive test battery. Future studies could counterbalance or randomize the test order where feasible, or select only individual cognitive tests relevant to the specific research question. This may also reduce or distribute fatigue effects across cognitive tests.

Furthermore, the duration of the cognitive assessments potentially led to increased cognitive fatigue, as indicated by a trend toward lower restfulness levels after the cognitive battery. This may have affected task performance, particularly in the later cognitive tasks [110]. Notably, changes in restfulness levels before and after completing the cognitive test battery were highly individual, and some participants may be more prone to cognitive fatigue than others, including those with OSA [111]. Cognitive assessments in OSA should therefore be administered with careful consideration of the assessment duration, test order, and opportunities for short breaks.

In addition, the length and structure of the cognitive battery affected the timing of the delayed recalls in the AVLT and VDLT, resulting in more variable and shorter recall intervals than the traditional delay [75,112]. Participants varied greatly in the time required to read instructions, finish practice trials, and complete cognitive tasks without time limits due to differences in response times (ITT, AVLT, ToL, both CBTTs, VDLT, IGT, ER-40, JLAP-15), performance (both CBTTs), or the option to quit early (PASAT level 3, data not analyzed). This complicated the standardization of recall intervals. In addition, test-order effects had to be taken into account. Thus, relatively short delayed recall intervals were chosen to limit variability across participants and to avoid assessment breaks, as the cognitive battery and the overall duration of the university visit required a reasonable amount of time. These shorter and more variable delayed recall intervals may have influenced delayed recall performance and therewith psychometric properties. Notably, shorter delayed recall intervals may be sufficient in some clinical populations [113,114], but have not yet been studied in OSA populations. Future studies should therefore investigate the optimal recall interval for individuals with OSA and aim for a standardized approach.

Another limitation of the cognitive test battery was the lack of alternate forms, which have been shown to reduce practice effects [22]. However, while alternate forms of cognitive tests can reduce practice effects stemming from memory for individual test-items, practice effects from test-taking strategies may remain [23,63]. Additionally, the currently available alternate forms are not always truly equivalent in test difficulty, which can result in higher or lower test scores at retest and affect score interpretation [20,22]. Therefore, there is an urgent need for novel and more alternate forms [67].

Furthermore, not all cognitive tasks included practice trials, which may have led to inconsistencies in performance across tasks. Although the cognitive battery included detailed instructions for all tasks, additional explanations and practice trials, which must be solved correctly to proceed to the test phase, could have been useful to support understanding, especially for older participants [103].

Lastly, the cognitive assessments were conducted only twice, with varying retest intervals between participants, which may have affected practice effects and test–retest reliability. Practice effects differ between the number of assessments and are especially prominent in the first few administrations before reaching a plateau, regardless of age, education, or intelligence [79,115]. Cognitive tests measuring executive functions

and processing speed seem particularly susceptible to strong practice effects early on [79,115], as also seen in our results. It is therefore likely that the cognitive tasks measuring executive functions and processing speed in this study will produce fewer practice effects and plateau in scores after multiple administrations. Furthermore, although practice effects were not associated with the length of the retest interval in our study, others have found that shorter intervals increase practice effects, whereas longer intervals decrease them [35,115]. Simultaneously, longer intervals lower test–retest reliability due to true changes in cognitive performance [22]. It is therefore crucial to examine how the number of repeated cognitive assessments and varying retest intervals affect psychometric properties. This can help to determine the optimal retest interval and interpret performance scores, considering that follow-up periods in OSA intervention studies can range from weeks to years [116–118]. Future studies should also examine a variety of cognitive tests to determine the most robust and clinically meaningful tests for all cognitive domains in OSA, use more difficult items and alternative scoring methods, and evaluate psychometric properties in clinical OSA cohorts.

4.3. General recommendations

While this study focused on a specific selection of cognitive tasks and their suitability for repeated assessments, we encourage extending the psychometric assessment to a variety of cognitive tests and to multiple cognitive domains, including attention, perceptual-motor skills, memory and learning, executive functions, and social cognition. Therefore, we propose the following general recommendations and encourage further research to establish a standardized cognitive test battery for measuring disorder progression and treatment effects in individuals with OSA.

1. Select cognitive test parameters with sufficient psychometric properties for repeated assessments, not necessarily their ability to distinguish between individuals with OSA and healthy controls.
2. Evaluate psychometric properties in individuals with OSA, and preferably across OSA subgroups, instead of relying on results from healthy populations.
3. Minimize test-order effects wherever possible, and consider the effect of using individual tests vs. a cognitive battery.
4. Include practice trials to ensure understanding of test instructions.
5. Consider the effect of different input or output modalities, particularly for tests relying heavily on graphomotor functions.
6. Minimize and account for practice effects by using alternate or parallel test versions, establishing a baseline over multiple assessments where advised, and by choosing an appropriate retest interval.
7. Reduce floor and ceiling effects by considering test difficulty, score range, and alternate scoring methods.
8. Further optimize task parameters, including the number of trials, stimulus and response types, and scoring methods.

4.4. Conclusion

This is the first study to assess the test–retest reliability, practice effects, and floor and ceiling effects of a comprehensive digital cognitive battery to determine its suitability for repeated measurements in individuals with OSA. The psychometric properties varied substantially across cognitive tasks and domains, underscoring the importance of careful test selection and the need for modifications to test design and scoring methods in longitudinal assessments.

Overall, our findings mark a step towards reliable, standardized cognitive assessments and provide guidance on selecting cognitive tests to monitor changes in cognitive performance related to disorder progression and treatment effects. Further research is needed to establish consensus on the optimal administration of cognitive tests in OSA, how to refine cognitive tests for different OSA subgroups, and the validity and clinical utility of digital cognitive assessments.

CRediT authorship contribution statement

Elena Richert: Writing – review & editing, Writing – original draft, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Minna Pitkänen:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **María K. Jónsdóttir:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Conceptualization. **Erna Sif Arnardóttir:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Funding acquisition. **Samu Kainulainen:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Timo Leppänen:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Juha Töyräs:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Henri Korkalainen:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Kamilla Rún Jóhannsdóttir:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Ethics declaration

Written informed consent to take part in the study and to publish the article has been obtained from all participants or their legal representatives. The privacy rights of participants have been observed.

This study was performed in compliance with relevant laws, regulatory frameworks and guidelines where the research took place. This study was approved by the National Bioethics Committee and Data Protection Authority of Iceland. (Approval No. VSN number 21–070)

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT in order to rephrase individual sentences and Grammarly to improve clarity, readability, and grammatical accuracy. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Sleep Revolution reports financial support was provided by Horizon 2020. Samu Kainulainen, Timo Leppänen, Minna Pitkänen reports financial support was provided by Research Council of Finland. Samu Kainulainen, Timo Leppänen, Minna Pitkänen, Henri Korkalainen reports financial support was provided by State Research Funding for university-level health research, Kuopio University Hospital, Wellbeing Service County of North Savo. Minna Pitkänen, Juha Töyräs reports financial support was provided by Sigrid Jusélius Foundation. Samu Kainulainen reports was provided by Jane and Aatos Erkko Fundation. Erna Sif Arnardóttir reports a relationship with Nox Medical that includes: speaking and lecture fees. Erna Sif Arnardóttir reports a relationship with ResMed that includes: speaking and lecture fees. Erna Sif Arnardóttir reports a relationship with Jazz Pharmaceuticals that includes: speaking and lecture fees. Erna Sif Arnardóttir reports a

relationship with Linde Healthcare that includes: speaking and lecture fees. Erna Sif Arnardóttir reports a relationship with Wink Sleep that includes: speaking and lecture fees. Erna Sif Arnardóttir reports a relationship with Apnimed that includes: speaking and lecture fees. Erna Sif Arnardóttir reports a relationship with Vistor that includes: speaking and lecture fees. Erna Sif Arnardóttir reports a relationship with Philips that includes: board membership. Erna Sif Arnardóttir reports a relationship with Lilly that includes: board membership. Erna Sif Arnardóttir reports a relationship with Sleep Cycle that includes: board membership. Erna Sif Arnardóttir is currently President of the European Sleep Research Society. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was carried out as a part of the Sleep Revolution project, which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 965417. SK, MP and TL are supported by the Research Council of Finland (grant numbers 360119 and 361199). SK, MP, TL, and HK are supported by the State Research Funding for university-level health research, Kuopio University Hospital, Wellbeing Service County of North Savo (project numbers 5041803, 5041804, 5041820, and 5041822). Additionally, MP and JT are supported by the Sigrid Jusélius Foundation, and SK by the Jane & Aatos Erkko Foundation. ESA discloses lecture fees from Nox Medical, ResMed, Jazz Pharmaceuticals, Linde Healthcare, Wink Sleep, Apnimed and Vistor. ESA is a previous member of the Philips Sleep Medicine & Innovation Medical Advisory Board and the Lilly Medical Advisory Board and is currently on the Sleep Cycle Data Monitoring Committee, and the President of the European Sleep Research Society. We are grateful to all participants for their time and commitment to this study. We also thank all staff and students of the Reykjavik University Sleep Institute for their support with data collection, sleep scorings, and data management.

Appendix A. Detailed cognitive task information

A.1. Simple Reaction Time Task (SRTT)

In the SRTT [119], participants were presented with a black fixation cross, followed by a target stimulus (a red circle). The target was presented at intervals of 2000 ms to 8000 ms in 1000 ms increments with a total of 40 trials. The mean reaction time over all 40 trials was calculated.

A.2. Inspection Time Task (ITT)

The ITT was modeled after Deary et al. [120]. The participants were presented with a simple visual stimulus consisting of two parallel vertical lines (arms), connected by a horizontal line at the top. One arm was long, the other one short, with the short arm being half the length of the long arm. The stimuli appeared for variable durations of 19, 25, 31, 37, 44, 50, 62, 75, 87, 100, 125, 150, or 200 ms, followed by a backward mask for 500 ms, where both arms were the same length. The participants were instructed to determine whether the short arm was on the left or right side by pressing the corresponding response key and were informed that reaction time was not important. The interstimulus interval (ISI) was 500 ms from response to cue onset, with the cue (a plus sign) lasting 500 ms. The task began with a practice block consisting of 25 trials with randomly selected stimulus durations and included visual feedback for correct and incorrect responses. This was followed by a test block with 75 trials in which each stimulus duration was presented on average six times (range 4–8) and no feedback was provided. Our analysis focused on the average proportion of correct responses across all stimulus durations.

A.3. Auditory Verbal Learning Task (AVLT)

The AVLT was modeled after the Rey Auditory Verbal Learning Test [112]. The participants were instructed to listen to and remember a list of 15 Icelandic words [121] prerecorded by a native speaker. The words were presented in a fixed order, one word at a time, at one word per second. After presenting all the words, the participants typed as many words as they could remember from the list into a text box. This sequence of list presentation and word recall was repeated five times. Additionally, a delayed recall took place after the completion of the Psychomotor Vigilance Task (PVT) and the Symbol Digit Modalities Task (SDMT), total of approximately 13–15 min). Minor errors, such as typographical mistakes or orthographic variations, such as the substitution of d for ð in Icelandic, were manually reviewed. In this paper, we analyzed the total recall score (sum of all correctly recalled items across all five immediate recalls), the delayed recall score (number of correctly recalled items during delayed recall), and the mean number of intrusions (entered words that were not part of the presented list) across all trials.

A.4. Psychomotor Vigilance Task (PVT)

In the PVT, participants were instructed to press the spacebar as soon as the target appeared, in the form of a red stopwatch with white digits (seconds: milliseconds) as described in Thomann et al. [122]. The task began with a habituation phase of 1 min, during which no responses were recorded, and seamlessly transitioned into the 10 min test phase. The ISI was between 2–10 s. Responses were captured relative to target onset and participants received feedback on their reaction times (ms), which was displayed for 1000 ms. If no response was given within 30 s after target onset, or if a response was recorded before target onset, the messages "OVERRUN!" or "Too early!", respectively, were displayed for 300 ms. In this study, we compared the overall mean reaction time, the reaction times for responses $\leq$ 500 ms, and the number of lapses, defined as responses slower than 500 ms.

A.5. Symbol Digit Modalities Task (SDMT)

This task was modeled after the SDMT by Smith [123]. A key with the digits 1–9 was presented at the top of the screen, where each digit was paired with a simple symbol. Below the key, several rows of those symbols with blank boxes underneath were presented. Participants were instructed to type the corresponding digit for each symbol as quickly as possible. The task began with the presentation of nine training boxes, consisting of three pre-filled example boxes and six practice boxes, for which visual feedback was provided on correct or incorrect entries. Once all training boxes were filled correctly, the participants pressed a key to move to the test phase, which lasted 120 s without feedback or the option to correct a prior entry. The total count of correct entries was used in the statistical analysis.

A.6. Tower of London (ToL)

The ToL was first developed by Shallice [124]. Three colored disks were presented on three pegs of varying lengths. The first peg could hold all three disks, the second peg two disks, and the third peg one disk. The participants' task was to arrange the disks from an initial state to a pictured target state in the fewest possible moves, which were displayed on the screen. Disks could only be moved one at a time and placed on pegs with sufficient space. In this study, the task started with a practice trial, followed by twelve test trials. There was no time limit. The test trials were presented in a sequentially fixed order consisting of two trials with two moves, two trials with three moves, four trials with four moves, and four trials with five moves. The participants had three attempts to solve each trial within the allocated moves. Solving the trial on the first attempt resulted in three points, on the second trial in

two points, and on the third trial in one point, resulting in a maximum score of 36 points. After failing to solve the trial in three attempts, this trial resulted in a score of zero points and continued with the next trial in the sequence. Furthermore, participants could reset the current trial by clicking the 'reset' button or skip to the next trial by pressing the 'next' button. The total score, mean solution time of correctly solved items, and mean first-move latency (planning time) were included in the statistical tests.

A.7. Corsi Block Tapping Task (CBTT) — Forward version

The CBTT was modeled after Kessels et al. [125]. In this task, participants were presented with nine blue boxes that lit up in yellow in a fixed sequence. Participants were asked to repeat the sequence in the same order by clicking on the corresponding boxes. The task began with a sequence length of two and increased in increments of one up to nine. Each sequence length was presented twice. If at least one of the two sequences was entered correctly, the next sequence length was presented. If both trials in a sequence length were entered incorrectly, the task ended. For the statistical analysis, we used the longest correctly recalled sequence length (max = 9).

A.8. Corsi Block Tapping Task (CBTT) — Backwards version

This task was similar to the CBTT forward version, except that the sequence had to be recalled in reverse order in the backwards version, with the longest possible sequence being eight instead of nine.

A.9. Visual Design Learning Task (VDLT)

The VDLT was modeled after the Rey Visual Design Learning Test [75]. The participants were presented with 15 simple figures, each displayed for 2 s, one at a time. After all figures had been displayed, the immediate recall followed, in which participants were instructed to draw all figures they could remember on the computer screen using the computer mouse. This was repeated three times. A delayed recall took place after the Iowa Gambling Task (approximately 4–6 min), in which the participants drew the 15 figures from memory, followed by a recognition task. In the recognition task, 15 previously presented figures and 15 new figures were displayed in a 5 × 6 matrix, and participants were instructed to select all the figures they remembered. Once a figure was selected, it got erased from the matrix. The task ended when the participants had selected 15 figures. In this analysis, we included only the hit rate for correctly selected figures in the recognition task. We omitted recall performance as it was not always possible to reliably identify which target figure a participant intended to draw, particularly with visually similar figures.

A.10. Iowa Gambling Task (IGT)

The IGT was modeled after Bechara et al. [126]. The participants were instructed to draw 100 cards from four card decks, with each card resulting in a win or a loss of money. Two of the card decks were overall advantageous (smaller rewards, smaller losses), and two were overall disadvantageous (large rewards, but also occasional large losses). The participants started with $2000, and the goal was to maximize the profit. The amount of money at the end was included in the statistical testing.

A.11. Penn emotional recognition task (ER-40)

The ER-40 was developed by Carter et al. [127]. This task included 40 faces in total, presented in randomized order, one face at a time. The participant's task was to select the corresponding emotion (anger, fear, happiness, sadness, or neutral). The task began with one practice trial, during which feedback was provided. Once the practice trial was answered correctly, the task continued with 40 test trials. The faces were balanced for age, ethnicity, gender, emotion, and intensity (four high-intensity, four low-intensity). We included the proportion of correct responses in the statistical analysis.

A.12. Judgment of Line Angle and Position (JLAP-15)

This task was modeled after Collaer et al. [128] and kept the original item presentation order. In this task, the participants were presented with a line at a specific angle, and instructed to identify the matching line from 15 options below. All lines were of the same length and equally distributed over a 180-degree half-circle. A total of 20 lines were presented one at a time. The time limit was set to 10 s per line, indicated by a timer in the corner. The total number of correct selections was included in the statistical testing.

A.13. Task Switching Paradigm (TSP)

The TSP was modeled after Rogers and Monsell [129] and included two categorization tasks: classifying a letter as a consonant or a vowel, and classifying a digit as odd or even. The same two response keys were used for consonants and even numbers, and for vowels and odd numbers. The presented stimulus consisted of a task-dependent target character (letter or digit) and a distractor character, e.g. '2F' in the letter task and '9#' in the digit task. The task began with four practice blocks, alternating between the letter and digit task, each with 24 trials. During practice, participants received feedback on incorrect responses and could only proceed once they had corrected the response. There was no time limit during the practice phase. During the test phase, a 2 × 2 matrix was presented, and the response timeout was set to 5000 ms. The stimuli rotated clockwise through the four matrix boxes, indicating the corresponding task through their location within the matrix. The top two boxes required the letter task, and the bottom boxes required the digit task, resulting in a task switch on every odd trial and no task switch on every even trial. The participants were instructed to respond as quickly and accurately as possible and were allowed to take short breaks between blocks. This test phase consisted of two crosstalk (C) and two non-crosstalk (N) blocks presented in the order CCNN. Each block started with twelve warm-up trials (not included in results) and was followed by 36 test trials (18 letter task trials, 18 digit task trials). In the C blocks, distractor characters belonged to three groups, including control (neutral non-alphanumeric distractor character), congruent (distractor character from the opposite task requiring the same key response as the target character), and incongruent (distractor character belongs to the opposite task and requires the opposite key response), with equal group proportions across trials. In the *N* blocks, distractor characters always belonged to the control group. For this paper, we analyzed the latency switch costs (difference between the mean latency of switch-trials and non-switch-trials, positive values indicate participants were slower on switch trials) of correct responses separately for C and *N* trials, and the accuracy switch cost (difference between proportion of correct switch trials and non-switch trials, negative values indicate lower accuracy on switch trials) separately for C and *N* trials. Trials with no response or latencies below 100 ms were excluded.

Table B.1
Results of the multiple linear regression test.

	Age		Time interval		Overall model
Test variable	Estimate (standard error)	*p*-value	Estimate (standard error)	*p*-value	*p*-value
SRTT, mean reaction time	−0.000 (0.013)	0.972	−0.012 (0.011)	0.277	0.542
ITT, proportion of correct	0.016 (0.012)	0.187	−0.011 (0.013)	0.419	0.224
AVLT, recall score	−0.002 (0.011)	0.855	−0.010 (0.010)	0.315	0.594
AVLT, delayed recall score	0.014 (0.010)	0.201	−0.007 (0.010)	0.453	0.321
AVLT, mean intrusions	0.004 (0.010)	0.734	−0.004 (0.010)	0.724	0.884
PVT, mean reaction time	−0.017 (0.011)	0.116	−0.007 (0.013)	0.568	0.271
PVT, mean reaction time ≤ 500 ms	−0.016 (0.010)	0.109	0.004 (0.011)	0.722	0.249
PVT, lapses	0.007 (0.012)	0.538	0.013 (0.014)	0.346	0.582
SDMT, number of correct entries	−0.002 (0.010)	0.848	−0.010 (0.010)	0.329	0.608
ToL, mean solution time	0.004 (0.010)	0.722	−0.006 (0.010)	0.523	0.754
ToL, mean first-move latency	0.001 (0.010)	0.926	−0.010 (0.010)	0.320	0.600
ToL, total score	0.021 (0.011)	0.074	−0.002 (0.010)	0.866	0.185
CBTT, forward, block span	−0.001 (0.012)	0.910	0.000 (0.010)	0.994	0.994
CBTT, backward, block span	0.005 (0.011)	0.644	0.013 (0.010)	0.194	0.396
VDLT, hit rate	0.012 (0.011)	0.273	−0.013 (0.010)	0.213	0.212
IGT, amount of coins won	−0.013 (0.010)	0.190	0.006 (0.010)	0.506	0.324
ER-40, proportion of correct	−0.012 (0.011)	0.907	0.007 (0.010)	0.505	0.791
JLAP-15, total correct	0.004 (0.010)	0.661	0.009 (0.010)	0.364	0.610
TSP, ACC, switch cost C	−0.006 (0.011)	0.576	0.007 (0.010)	0.462	0.637
TSP, RT, switch cost C	0.009 (0.011)	0.461	−0.007 (0.010)	0.466	0.572
TSP, ACC, switch cost N	0.009 (0.012)	0.458	0.012 (0.010)	0.232	0.387
TSP, RT, switch cost N	0.006 (0.010)	0.546	−0.015 (0.009)	0.116	0.233
PASAT1, number of correct	0.001 (0.010)	0.951	−0.004 (0.010)	0.680	0.914
PASAT2, number of correct	−0.001 (0.011)	0.951	0.005 (0.010)	0.658	0.904

ACC = Accuracy, AVLT = Auditory verbal learning task, CBTT = Corsi block tapping task, ER-40 = Penn emotional recognition task, IGT = Iowa gambling task, ITT = Inspection time task, JLAP-15 = Judgment of line angle and position, PASAT1 = Paced auditory serial addition task level 1, PASAT2 = Paced auditory serial addition task level 2, PVT = Psychomotor vigilance task, RT = Reaction time, SDMT = Symbol digit modalities task, SRTT = Simple reaction time task, ToL = Tower of London, TSP = Task switching paradigm, VDLT = Visual design learning task.

A.14. Paced Auditory Serial Addition Task (PASAT)

The PASAT was first introduced by Gronwall [130]. In this version, the participants listened to digits (1–9) recorded by a native speaker, one digit at a time, in randomized order. After each new digit was presented, the participant's task was to add the current digit to the previous one and indicate the correct sum by clicking the corresponding box on the screen, with numbers 1–18 arranged in a full circle. The task began with a practice round, which included 11 digits presented at an ISI of 4 s and provided visual feedback about correct or incorrect answers. Furthermore, all rounds provided auditory feedback in the form of an aversive noise burst played for 100 ms when a response was incorrect. The practice round was followed by level 1 with a total of 60 digits and an ISI of 3 s, and level 2 with 150 digits and an ISI of 2 s. Level 3 was presented at an ISI of 1.5 s and included a maximum of 400 digits and a 'quit' button to end the task at any time. We included the number of correct responses in level 1 and level 2 for analysis. Although level 3 was presented, the number of correct responses was close to 0 for almost all participants, either due to the fast digit presentation or the option to quit the task during level 3 and was therefore omitted from analysis.

Appendix B. Multiple regression results

See Table B.1.

References

[1] Gosselin N, Baril A-A, Osorio RS, Kaminska M, Carrier J. Obstructive sleep apnea and the risk of cognitive decline in older adults. Am J Respir Crit Care Med 2019;199(2):142–8.

[2] Heilbronner RL, Sweet JJ, Attix DK, Krull KR, Henry GK, Hart RP. Official position of the American academy of clinical neuropsychology on serial neuropsychological assessments: the utility and challenges of repeat test administrations in clinical and forensic contexts. Clin Neuropsychol 2010;24(8):1267–78.

[3] Bubu OM, Andrade AG, Umasabor-Bubu OQ, Hogan MM, Turner AD, de Leon MJ, Ogedegbe G, Ayappa I, Jackson ML, Varga AW, et al. Obstructive sleep apnea, cognition and alzheimer's disease: a systematic review integrating three decades of multidisciplinary research. Sleep Med Rev 2020;50:101250.

[4] Jóhannsdóttir KR, Ferretti D, Árnadóttir BS, Jónsdóttir MK. Objective measures of cognitive performance in sleep disorder research. Sleep Med Clin 2021;16(4):575–93, Publisher: Elsevier.

[5] Kainulainen S, Duce B, Korkalainen H, Oksenberg A, Leino A, Arnardóttir ES, Kulkas A, Myllymaa S, Töyräs J, Leppänen T. Severe desaturations increase psychomotor vigilance task-based median reaction time and number of lapses in obstructive sleep apnoea patients. Eur Respir J 2020;55(4).

[6] Pahari P, Korkalainen H, Karhu T, Rissanen M, Arnardóttir ES, Hrubos-Strøm H, Duce B, Töyräs J, Leppänen T, Nikkonen S. Obstructive sleep apnea-related intermittent hypoxaemia is associated with impaired vigilance. J Sleep Res 2023;32(3):e13803.

[7] Staykov E, Mann DL, Leppänen T, Töyräs J, Kainulainen S, Azarbarzin A, Duce B, Sands SA, Terrill PI, Group SCI. Increased flow limitation during sleep is associated with decreased psychomotor vigilance task performance in individuals with suspected obstructive sleep apnea: a multi-cohort study. Sleep 2024;47(6):zsae077.

[8] Batool-Anwar S, Kales SN, Patel SR, Varvarigou V, DeYoung PN, Malhotra A. Obstructive sleep apnea and psychomotor vigilance task performance. Nat Sci Sleep 2014;6:65–71.

[9] Lee I-S, Bardwell WA, Ancoli-Israel S, Dimsdale JE. Number of lapses during the psychomotor vigilance task as an objective measure of fatigue. J Clin Sleep Med 2010;6(2):163–8.

[10] Thorisdottir K, Hrubos-Strøm H, Karhu T, Nikkonen S, Dammen T, Nordhus IH, Leppänen T, Jónsdóttir MK, Arnardóttir ES. Verbal memory is linked to average oxygen saturation during sleep, not the apnea-hypopnea index nor novel hypoxic load variables. Sleep Med 2024;123:29–36.

[11] Twigg GL, Papaioannou I, Jackson M, Ghiassi R, Shaikh Z, Jaye J, Graham KS, Simonds AK, Morrell MJ. Obstructive sleep apnea syndrome is associated with deficits in verbal but not visual memory. Am J Respir Crit Care Med 2010;182(1):98–103.

[12] Ayalon L, Ancoli-Israel S, Klemfuss Z, Shalauta MD, Drummond SP. Increased brain activation during verbal learning in obstructive sleep apnea. Neuroimage 2006;31(4):1817–25.

[13] Holloway BM, Harding CD, DeYoung P, Kwan CG, Avetisyan L, Lui KK, Ancoli-Israel S, Banks SJ, Djonlagic I, Malhotra A. Comorbid insomnia and sleep apnea is associated with worse verbal episodic memory in older females. J Clin Sleep Med 2025;21(12):2129–38.

[14] Naëgelé B, Launois SH, Mazza S, Feuerstein C, Pépin J-L, Lévy P. Which memory processes are affected in patients with obstructive sleep apnea? An evaluation of 3 types of memory. Sleep 2006;29(4):533–44.

[15] Salorio CF, White DA, Piccirillo J, Duntley SP, Uhles ML. Learning, memory, and executive control in individuals with obstructive sleep apnea syndrome. J Clin Exp Neuropsychol 2002;24(1):93–100.

[16] Tulek B, Atalay NB, Kanat F, Suerdem M. Attentional control is partially impaired in obstructive sleep apnea syndrome. J Sleep Res 2013;22(4):422–9.

[17] Mathieu A, Mazza S, Décary A, Massicotte-Marquez J, Petit D, Gosselin N, Malo J, Montplaisir J. Effects of obstructive sleep apnea on cognitive function: a comparison between younger and older OSAS patients. Sleep Med 2008;9(2):112–20.

[18] Daurat A, Ricarrère M, Tiberge M. Decision making is affected in obstructive sleep apnoea syndrome. J Neuropsychol 2013;7(1):139–44.

[19] Ferini-Strambi L, Baietto C, Di Gioia M, Castaldi P, Castronovo C, Zucconi M, Cappa S. Cognitive dysfunction in patients with obstructive sleep apnea (OSA): partial reversibility after continuous positive airway pressure (CPAP). Brain Res Bull 2003;61(1):87–92.

[20] Calamia M, Markon K, Tranel D. The robust reliability of neuropsychological measures: Meta-analyses of test–retest correlations. Clin Neuropsychol 2013;27(7):1077–105.

[21] Salthouse TA. Robust cognitive change. J Int Neuropsychol Soc 2012;18(4):749–56.

[22] Calamia M, Markon K, Tranel D. Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. Clin Neuropsychol 2012;26(4):543–70.

[23] Duff K. Evidence-based indicators of neuropsychological change in the individual patient: relevant concepts and methods. Arch Clin Neuropsychol 2012;27(3):248–61.

[24] Lo AH, Humphreys M, Byrne GJ, Pachana NA. Test–retest reliability and practice effects of the wechsler memory scale-III. J Neuropsychol 2012;6(2):212–31.

[25] Hinton-Bayre AD. Clarifying discrepancies in responsiveness between reliable change indices. Arch Clin Neuropsychol 2016;31(7):754–68.

[26] Goldberg TE, Goldman RS, Burdick KE, Malhotra AK, Lencz T, Patel RC, Woerner MG, Schooler NR, Kane JM, Robinson DG. Cognitive improvement after treatment with second-generation antipsychotic medications in first-episode schizophrenia: is it a practice effect? Arch Gen Psychiatry 2007;64(10):1115–22.

[27] Jutten RJ, Grandoit E, Foldi NS, Sikkes SA, Jones RN, Choi S-E, Lamar ML, Louden DK, Rich J, Tommet D, et al. Lower practice effects as a marker of cognitive performance and dementia risk: a literature review. Alzheimer's Dement: Diagn Assess Dis Monit 2020;12(1):e12055.

[28] Wang L, Zhang Z, McArdle JJ, Salthouse TA. Investigating ceiling effects in longitudinal data analysis. Multivar Behav Res 2008;43(3):476–96.

[29] Lewis-Beck MS, Bryman A, Liao TF, editors. Encyclopedia entry: Floor effect. In: The SAGE Encyclopedia of Social Science Research Methods. Thousand Oaks, CA: SAGE; 2004.

[30] Uttl B. Measurement of individual differences: lessons from memory assessment in research and clinical practice. Psychol Sci 2005;16(6):460–7.

[31] Liu Q, Wang L. t-Test and ANOVA for data with ceiling and/or floor effects. Behav Res Methods 2021;53(1):264–77.

[32] Hammers DB, Bothra S, Polsinelli A, Apostolova LG, Duff K. Evaluating practice effects across learning trials–ceiling effects or something more? J Clin Exp Neuropsychol 2024;46(7):630–43.

[33] Fernández-Marcos T, de la Fuente C, Santacreu J. Test–retest reliability and convergent validity of attention measures. Appl Neuropsychol: Adult 2018;25(5):464–72.

[34] Germine L, Reinecke K, Chaytor NS. Digital neuropsychology: Challenges and opportunities at the intersection of science and software. Clin Neuropsychol 2019;33(2):271–86.

[35] Bergman I, Franke Föyen L, Gustavsson A, Van den Hurk W. Test–retest reliability, practice effects and estimates of change: A study on the mindmore digital cognitive assessment tool. Scand J Psychol 2025;66(1):1–14.

[36] Zygouris S, Tsolaki M. Computerized cognitive testing for older adults: a review. Am J Alzheimer's Dis Other Dementias® 2015;30(1):13–28.

[37] Cubillos C, Rienzo A, et al. Digital cognitive assessment tests for older adults: systematic literature review. JMIR Ment Health 2023;10(1):e47487.

[38] Holm SP, Wolfer AM, Pointeau GH, Lipsmeier F, Lindemann M. Practice effects in performance outcome measures in patients living with neurologic disorders–A systematic review. Heliyon 2022;8(8).

[39] Charalambous AP, Pye A, Yeung WK, Leroi I, Neil M, Thodi C, Dawes P. Tools for app-and web-based self-testing of cognitive impairment: systematic search and evaluation. J Med Internet Res 2020;22(1):e14551.

[40] Ferretti D, Richert E, Zou D, Hedner J, Strassberger C, Arnardóttir ES, Grote L, Jóhannsdóttir KR. Impaired emotion recognition is associated with sleep-related hypoxic burden. Sleep 2026;49(7):zsag078.

[41] Thorisdottir K, Leppänen T, Hrubos-Strøm H, Richert E, Biedebach L, Jóhannsdóttir KR, Arnardóttir ES, Jónsdóttir MK. Hypoxic load and average oxygen-saturation during sleep and wake are associated with cognitive function in obstructive sleep apnea. J Sleep Res 2026;35(1):e70136.

[42] Arnardóttir ES, Islind AS, Óskarsdóttir M, Ólafsdóttir KA, August E, Jónasdóttir L, Hrubos-Strøm H, Saavedra JM, Grote L, Hedner J, Höskuldsson S, Ágústsson JS, Jóhannsdóttir KR, McNicholas WT, Pevernagie D, Sund R, Töyräs J, Leppänen T, on behalf of the Sleep Revolution. The Sleep Revolution project: the concept and objectives. J Sleep Res 2022;31(4):e13630.

[43] Sveinbjarnarson BF, Schmitz L, Arnardóttir ES, Islind AS. The sleep revolution platform: A dynamic data source pipeline and digital platform architecture for complex sleep data. Curr Sleep Med Rep 2023;9(2):91–100.

[44] Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009;42(2):377–81.

[45] Johns M, Hocking B. Daytime sleepiness and sleep habits of Australian workers. Sleep 1997;20(10):844–7.

[46] Montalvão Neves J, Maciel CA, Alexandre-Sousa P. Benefits of treating obstructive sleep apnea in cognition: systematic literature review. J Dement Alzheimer's Dis 2025;2(1):3.

[47] Xu S, Akioma M, Yuan Z. Relationship between circadian rhythm and brain cognitive functions. Front Optoelectron 2021;14(3):278–87.

[48] Inquisit (Version 651). [Computer software]. 2021, Retrieved from https://www.millisecond.com.

[49] Ferretti D, Islind AS, Ólafsdóttir KA, Sigurdardottir S, Jóhannsdóttir KR, Hedner J, Grote L, Arnardóttir ES. Feasibility and usability of three consecutive nights with self-applied, home polysomnography. J Sleep Res 2025;34(2):e14286.

[50] Rusanen M, Korkalainen H, Gretarsdottir H, Siilak T, Olafsdottir KA, Töyräs J, Myllymaa S, Arnardóttir ES, Leppänen T, Kainulainen S. Self-applied somnography: technical feasibility of electroencephalography and electro-oculography signal characteristics in sleep staging of suspected sleep-disordered adults. J Sleep Res 2023;e13977, Publisher: Wiley Online Library.

[51] Kainulainen S, Korkalainen H, Sigurðardóttir S, Myllymaa S, Serwatko M, Sigurðardóttir SÞ, Clausen M, Leppänen T, Arnardóttir ES. Comparison of EEG signal characteristics between polysomnography and self applied somnography setup in a pediatric cohort. IEEE Access 2021;9:110916–26, Publisher: IEEE.

[52] Berry RB, Quan SF, Abreu AR, Bibbs ML, DelRosso L, Harding SM, Mao M-M, Plante DT, Pressman MR, Troester MM, Vaughn BV. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications (Version 2.6). American Academy of Sleep Medicine; 2020.

[53] AASM Task Force. Sleep-related breathing disorders in adults: recommendations for syndrome definition and measurement techniques in clinical research. The report of an American academy of sleep medicine task force. Sleep 1999;22:667–89.

[54] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2023.

[55] Gamer M, Lemon J, Singh IFP. irr: Various coefficients of interrater reliability and agreement. 2019, R package version 0.84.1.

[56] Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: Tests in linear mixed effects models. 2017, R package version 3.1.3.

[57] Benedict RH, DeLuca J, Phillips G, LaRocca N, Hudson LD, Rudick R, Consortium MSOA. Validity of the symbol digit modalities test as a cognition performance outcome measure for multiple sclerosis. Mult Scler J 2017;23(5):721–33.

[58] Strober L, Bruce J, Arnett P, Alschuler K, DeLuca J, Chiaravalloti N, Lebkuecher A, Di Benedetto M, Cozart J, Thelen J, et al. A much needed metric: Defining reliable and statistically meaningful change of the oral version symbol digit modalities test (SDMT). Mult Scler Relat Disord 2022;57:103405.

[59] Balkin TJ, Bliese PD, Belenky G, Sing H, Thorne DR, Thomas M, Redmond DP, Russo M, Wesensten NJ. Comparative utility of instruments for monitoring sleepiness-related performance decrements in the operational environment. J Sleep Res 2004;13(3):219–27.

[60] Langner R, Scharnowski F, Ionta S, Salmon CEG, Piper BJ, Pamplona GSP. Evaluation of the reliability and validity of computerized tests of attention. PLoS One 2023;18(1):e0281196.

[61] Dorrian J, Rogers NL, Dinges DF. Psychomotor vigilance performance: Neurocognitive assay sensitive to sleep loss. Sleep Deprivation 2004;39–70.

[62] Martínez-García MÁ, Labarca G. Obstructive sleep apnea in women: scientific evidence is urgently needed. J Clin Sleep Med 2022;18(1):1–2.

[63] Beglinger LJ, Gaydos B, Tangphao-Daniels O, Duff K, Kareken DA, Crawford J, Fastenau PS, Siemers ER. Practice effects and the use of alternate forms in serial neuropsychological testing. Arch Clin Neuropsychol 2005;20(4):517–29.

[64] Baird B, Tombaugh TN, Francis M. The effects of practice on speed of information processing using the adjusting-paced serial addition test (adjusting-PSAT) and the computerized tests of information processing (CTIP). Appl Neuropsychol 2007;14(2):88–100.

[65] Tombaugh TN. A comprehensive review of the paced auditory serial addition test (PASAT). Arch Clin Neuropsychol 2006;21(1):53–76.

[66] Rogers JM, Fox AM. Event-related potential practice effects on the paced auditory serial addition test (PASAT). Adv Cogn Psychol 2012;8(4):281.

[67] Lee Meeuw Kjoe PR, Vermeulen IE, Agelink van Rentergem JA, van der Wall E, Schagen S. Standardized item selection for alternate computerized versions of rey auditory verbal learning test (-based) word lists. J Clin Exp Neuropsychol 2022;44(9):681–701.
[68] Stricker NH, Christianson TJ, Lundt ES, Alden EC, Machulda MM, Fields JA, Kremers WK, Jack, Jr. CR, Knopman DS, Mielke MM, et al. Mayo normative studies: Regression-based normative data for the auditory verbal learning test for ages 30–91 years and the importance of adjusting for sex. J Int Neuropsychol Soc 2021;27(3):211–26.
[69] Thomas KR, Eppig J, Edmonds EC, Jacobs DM, Libon DJ, Au R, Salmon DP, Bondi MW. Word-list intrusion errors predict progression to mild cognitive impairment. Neuropsychology 2018;32(2):235.
[70] Raizner RD, Song J, Levin HS. Raising the ceiling: The tower of London-extended version. Dev Neuropsychol 2002;21(1):1–14.
[71] Unterrainer JM, Rahm B, Kaller CP, Wild PS, Münzel T, Blettner M, Lackner K, Pfeiffer N, Beutel ME. Assessing planning ability across the adult life span in a large population-representative sample: Reliability estimates and normative data for the tower of London (TOL-F) task. J Int Neuropsychol Soc 2019;25(5):520–9.
[72] Schnirman GM, Welsh MC, Retzlaff PD. Development of the tower of London-revised. Assessment 1998;5(4):355–60.
[73] Kaller CP, Unterrainer JM, Stahl C. Assessing planning ability with the tower of London task: psychometric properties of a structurally balanced problem set. Psychol Assess 2012;24(1):46.
[74] Schyle V, Schumacher LV, Rahm B, Unterrainer JM. Parallel and identical test–retest reliability of the tower of London test–freiburg version. J Int Neuropsychol Soc 2023;29(8):783–8.
[75] Wilhelm P. Reliability and validity of the rey visual design learning test in primary school children. J Clin Exp Neuropsychol 2004;26(7):981–94.
[76] Izgi B, Moore TM, Yalcinay-Inan M, Port AM, Kuscu K, Gur RC, Yapici Eser H. Test–retest reliability of the turkish translation of the penn computerized neurocognitive battery. Appl Neuropsychol: Adult 2022;29(5):1258–67.
[77] Lim K, Lee S-A, Pinkham AE, Lam M, Lee J. Evaluation of social cognitive measures in an Asian schizophrenia sample. Schizophr Res: Cogn 2020;20:100169.
[78] Blotenberg I, Schmidt-Atzert L. On the locus of the practice effect in sustained attention tests. J Intell 2019;7(2):12.
[79] White N, Flannery L, McClintock A, Machado L. Repeated computerized cognitive testing: Performance shifts and test–retest reliability in healthy older adults. J Clin Exp Neuropsychol 2019;41(2):179–91.
[80] Schmitz F, Kunina-Habenicht O, Hildebrandt A, Oberauer K, Wilhelm O. Psychometrics of the iowa and berlin gambling tasks: Unresolved issues with reliability and validity for risk taking. Assessment 2020;27(2):232–45.
[81] Sullivan-Toole H, Haines N, Dale K, Olino TM. Enhancing the psychometric properties of the iowa gambling task using full generative modeling. Comput Psychiatry 2022;6(1):189.
[82] Eisma YB, de Winter J. How do people perform an inspection time task? An examination of visual illusions, task experience, and blinking. J Cogn 2020;3(1):34.
[83] Bors DA, Stokes TL, Forrin B, Hodder SL. Inspection time and intelligence: Practice, strategies, and attention. Intelligence 1999;27(2):111–29.
[84] Vandierendonck A, Kemps E, Fastame MC, Szmalec A. Working memory components of the corsi blocks task. Br J Psychol 2004;95(1):57–79.
[85] Buelow MT, Suhr JA. Construct validity of the iowa gambling task. Neuropsychol Rev 2009;19(1):102–14.
[86] Van Patten R, Iverson GL, Muzeau MA, VanRavenhorst-Bell HA. Test–retest reliability and reliable change estimates for four mobile cognitive tests administered virtually in community-dwelling adults. Front Psychol 2021;12:734947.
[87] Collaer ML, Nelson JD. Large visuospatial sex difference in line judgment: Possible role of attentional factors. Brain Cogn 2002;49(1):1–12.
[88] Woods DL, Wyma JM, Herron TJ, Yund E. The effects of repeat testing, malingering, and traumatic brain injury on computerized measures of visuospatial memory span. Front Hum Neurosci 2016;9:690.
[89] Woods DL, Wyma JM, Yund EW, Herron TJ, Reed B. Factors influencing the latency of simple reaction time. Front Hum Neurosci 2015;9:131.
[90] Hughes MM, Linck JA, Bowles AR, Koeth JT, Bunting MF. Alternatives to switch-cost scoring in the task-switching paradigm: Their reliability and increased validity. Behav Res Methods 2014;46(3):702–21.
[91] Schmitz F, Krämer RJ. Task switching: on the relation of cognitive flexibility with cognitive capacity. J Intell 2023;11(4):68.
[92] Draheim C, Hicks KL, Engle RW. Combining reaction time and accuracy: The relationship between working memory capacity and task switching as a case example. Perspect Psychol Sci 2016;11(1):133–55.
[93] Wimms A, Woehrle H, Ketheeswaran S, Ramanan D, Armitstead J. Obstructive sleep apnea in women: specific issues and interventions. BioMed Res Int 2016;2016(1):1764837.
[94] Coates TJ, George JM, Killen JD, Marchini E, Hamilton S, Thorensen CE. First night effects in good sleepers and sleep-maintenance insomniacs when recorded at home. Sleep 1981;4(3):293–8.
[95] Ding L, Chen B, Dai Y, Li Y. A meta-analysis of the first-night effect in healthy individuals for the full age spectrum. Sleep Med 2022;89:159–65.
[96] Rechtschaffen A, Verdone P. Amount of dreaming: effect of incentive, adaptation to laboratory, and individual differences. Percept Mot Skills 1964;19(3):947–58.
[97] Watson D. Stability versus change, dependability versus error: Issues in the assessment of personality over time. J Res Pers 2004;38(4):319–50.
[98] Li Y, Lin S, Wei H, Chen Q, Chen S, Wei F, Li S, Qiu Y. Associations between excessive daytime sleepiness, cognitive impairment, and brain alterations in patients with obstructive sleep apnea. J Magn Reson Imaging 2026;63(6):1625–36.
[99] Zhou J, Camacho M, Tang X, Kushida CA. A review of neurocognitive function and obstructive sleep apnea with or without daytime sleepiness. Sleep Med 2016;23:99–108.
[100] Mortamais M, Ash JA, Harrison J, Kaye J, Kramer J, Randolph C, Pose C, Albala B, Ropacki M, Ritchie CW, et al. Detecting cognitive changes in preclinical alzheimer's disease: A review of its feasibility. Alzheimer's Dement 2017;13(4):468–92.
[101] Duff K, Lyketsos CG, Beglinger LJ, Chelune G, Moser DJ, Arndt S, Schultz SK, Paulsen JS, Petersen RC, McCaffrey RJ. Practice effects predict cognitive outcome in amnestic mild cognitive impairment. Am J Geriatr Psychiatry 2011;19(11):932–9.
[102] Avila Pérez S, Koppelmans V, Duff KM, Ruitenberg MF. One-year practice effects predict long-term cognitive outcomes in Parkinson's disease. J Parkinson's Dis 2025;15(4):858–67.
[103] Björngrim S, van den Hurk W, Betancort M, Machado A, Lindau M. Comparing traditional and digitized cognitive tests used in standard clinical evaluation–a study of the digital application minnemera. Front Psychol 2019;10:2327.
[104] Giaquinto F, Battista P, Angelelli P. Touchscreen cognitive tools for mild cognitive impairment and dementia used in primary care across diverse cultural and literacy populations: a systematic review. J Alzheimer's Dis 2022;90(4):1359–80.
[105] Gerth S, Dolk T, Klassert A, Fliesser M, Fischer MH, Nottbusch G, Festman J. Adapting to the surface: A comparison of handwriting measures when writing on a tablet computer and on paper. Hum Mov Sci 2016;48:62–73.
[106] Noad R, Newman C, Chynoweth J, Mayers J, Hall S, Murphy D. A pilot examination of the validity of stylus and finger drawing on visuomotor-mediated tests on acemobile. J Clin Exp Neuropsychol 2023;45(5):445–51.
[107] Lempke LB, Lynall RC, Anderson MN, McCrea MA, McAllister TW, Broglio SP, Schmidt JD. Optimizing order of administration for concussion baseline assessment among NCAA student-athletes and military cadets. Sports Med 2022;52(1):165–76.
[108] Lloyd KP, Higginson CI, Lating JM, Coiro MJ. Test-order effects: A measure of verbal memory influences performance on a measure of verbal fluency. Appl Neuropsychol: Adult 2012;19(4):299–304.
[109] Harvey PD. Domains of cognition and their assessment. Dialogues Clin Neurosci 2019;21(3):227–37.
[110] Holtzer R, Shuman M, Mahoney JR, Lipton R, Verghese J. Cognitive fatigue defined in the context of attention networks. Aging, Neuropsychol Cogn 2010;18(1):108–28.
[111] Benkirane O, Simor P, Mairesse O, Peigneux P. Sleep fragmentation modulates the neurophysiological correlates of cognitive fatigue. Clocks Sleep 2024;6(4):602–18.
[112] Schmidt M. Rey Auditory Verbal Learning Test. Western Psychological Services Los Angeles; 1996.
[113] Xu Y, Chen K, Zhao Q, Li F, Guo Q. Short-term delayed recall of auditory verbal learning test provides equivalent value to long-term delayed recall in predicting MCI clinical outcomes: a longitudinal follow-up study. Appl Neuropsychol: Adult 2020;27(1):73–81.
[114] Zhao Q, Lv Y, Zhou Y, Hong Z, Guo Q. Short-term delayed recall of auditory verbal learning test is equivalent to long-term delayed recall for identifying amnestic mild cognitive impairment. PLoS One 2012;7(12):e51157.
[115] Bartels C, Wegrzyn M, Wiedl A, Ackermann V, Ehrenreich H. Practice effects in healthy adults: a longitudinal study on frequent repetitive cognitive testing. BMC Neurosci 2010;11(1):118.
[116] Chaiard J, Bhatarasakoon P. Effectiveness of behavioral and psychosocial interventions for continuous positive airway pressure adherence among patients with obstructive sleep apnea: A systematic review and meta-analysis. Appl Nurs Res 2023;69:151654.
[117] Papageorgiou SN, Konstantinidis I, Papadopoulou AK, Apostolidou-Kiouti F, Avgerinos I, Pataka A, Eliades T, Tsapas A, Haidich A-B. Comparative efficacy of non-pharmacological interventions for adults with sleep apnea: A systematic review and network meta-analysis. Sleep Med 2025;128:130–8.
[118] Kang K-T, Yeh T-H, Ko J-Y, Lee C-H, Lin M-T, Hsu W-C. Effect of sleep surgery on blood pressure in adults with obstructive sleep apnea: a systematic review and meta-analysis. Sleep Med Rev 2022;62:101590.
[119] Donders FC. On the speed of mental processes (Koster WG, transl.). Acta Psychol 1969;30:412–31.
[120] Deary IJ, Simonotto E, Meyer M, Marshall A, Marshall I, Goddard N, Wardlaw JM. The functional anatomy of inspection time: an event-related fMRI study. Neuroimage 2004;22(4):1466–79.

[121] Þórisdóttir ÞB. Íslensk viðmiðunargildi fyrir Rey orðaminnispróf, Rey myndaminnispróf, Kennslaminnispróf Warringtons og Biber myndaminnispróf [Bachelor's thesis], Háskóli Íslands; 2001.
[122] Thomann J, Baumann CR, Landolt H-P, Werth E. Psychomotor vigilance task demonstrates impaired vigilance in disorders with excessive daytime sleepiness. J Clin Sleep Med 2014;10(9):1019–24.
[123] Smith A. Symbol Digit Modalities Test. Los Angeles: Western Psychological Services; 1973.
[124] Shallice T. Specific impairments of planning. Philos Trans R Soc Lond B, Biol Sci 1982;298(1089):199–209.
[125] Kessels RP, van den Berg E, Ruis C, Brands AM. The backward span of the corsi block-tapping task and its association with the WAIS-III digit span. Assessment 2008;15(4):426–34.
[126] Bechara A, Damasio AR, Damasio H, Anderson SW. Insensitivity to future consequences following damage to human prefrontal cortex. Cognition 1994;50:7–15.
[127] Carter CS, Barch DM, Gur R, Gur R, Pinkham A, Ochsner K. CNTRICS final task selection: social cognitive and affective neuroscience–based measures. Schizophr Bull 2009;35(1):153–62.
[128] Collaer ML, Reimers S, Manning JT. Visuospatial performance on an internet line judgment task and potential hormonal markers: sex, sexual orientation, and 2D: 4D. Arch Sex Behav 2007;36(2):177–92.
[129] Rogers RD, Monsell S. Costs of a predictible switch between simple cognitive tasks. J Exp Psychol [Gen] 1995;124(2):207.
[130] Gronwall D. Paced auditory serial-addition task: a measure of recovery from concussion. Percept Mot Skills 1977;44(2):367–73.