



# Remote digital cognitive assessment for aging and dementia using the Oxford Cognitive Testing Portal OCTAL



Sijia Zhao<sup>1,7</sup> ✉, Sofia Toniolo<sup>2,3,7</sup>, Qian-Yuan Tang<sup>4</sup>, Anna Scholcz<sup>1,3</sup>, Akke Ganse-Dumrath<sup>1,2</sup>, Claudia Gendarini<sup>2,5,6</sup>, M. John Broulidakis<sup>1,2</sup>, Sian Thompson<sup>2</sup>, Sanjay G. Manohar<sup>1,2,3</sup> & Masud Husain<sup>1,2,3</sup>

The global rise in dementia necessitates scalable cognitive assessments that can evolve to serve both clinical and research applications. We present the Oxford Cognitive Testing Portal (OCTAL), a remote, browser-based platform providing performance metrics for memory, attention, visuospatial and executive function domains. Four validation studies ( $N = 1664$ ) confirmed cross-cultural applicability, lifespan sensitivity and clinical utility. Task performance was equivalent in English- and Chinese-speaking younger adults and mapped domain-specific ageing trajectories in mid- to late-adulthood. In a memory-clinic cohort ( $N = 194$ ), 5-minute OCTAL screen distinguished patients with Alzheimer's disease dementia from subjective cognitive decline ( $AUC = 0.92$ ), matching a standard paper-based test, while a 20-minute subset surpassed this ( $AUC = 0.97$ ;  $p = 0.04$ ). Test-retest reliability was very good ( $ICC \geq 0.79$ ;  $N = 118$ ). OCTAL enables remote assessment for large-scale research and screening, with an open, modular architecture that makes it a uniquely sustainable and evolvable tool for the research community.

Dementia prevalence is increasing worldwide as global populations age, while capacity of memory assessment services has not kept pace<sup>1–3</sup>. In the United States alone, 7.2 million adults aged  $\geq 65$  are currently living with Alzheimer's disease (AD), a figure projected to almost double by 2060<sup>4</sup>. Meanwhile, workforce shortages in primary-care and specialist memory clinics, coupled with linguistic and cultural barriers, deprive many individuals of timely cognitive assessment and longitudinal monitoring<sup>3</sup>. There is also increasing recognition that screening for AD and other neurodegenerative conditions at scale might be important to introduce interventions—pharmacological and lifestyle—that might delay progression to dementia onset<sup>5</sup>.

Public demand for proactive cognitive evaluation is equally compelling, as reflected in a recent survey of 1702 US adults over 45, who were demographically matched to U.S. households<sup>4</sup>. The results showed that 99% view an early diagnosis of AD as important, while 59% believe routine cognitive screening is a vital part of preventive care<sup>4</sup>. These attitudes underscore both an ethical imperative and a growing demand for tests that (i) detect impairment with gold-standard accuracy, (ii) can be administered repeatedly without ceiling or practice effects, and (iii) scale via unsupervised, browser-based delivery acceptable to culturally diverse users. The need for

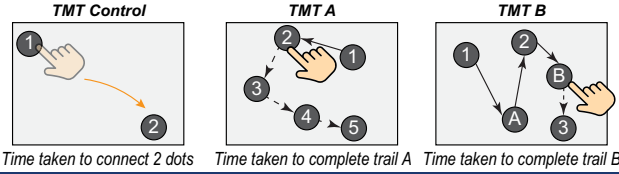
culturally adaptable approaches is particularly important given that most digital tools to date have been validated primarily in Western, educated, industrialised, rich and democratic populations<sup>6</sup>.

Yet, the current gold-standard cognitive screening tests — the Mini-Mental State Examination<sup>7</sup>, Montreal Cognitive Assessment<sup>8</sup> and Addenbrooke's Cognitive Examination-III (ACE-III)<sup>9</sup> — remain almost exclusively pen-and-paper. Each administration occupies 15–20 min of clinician time, limiting patient throughput in already overstretched services and, crucially, lengthening patients' waiting times<sup>10</sup>. Despite their clinical utility, their coarse scoring offers limited granularity for tracking subtle decline, particularly in individuals with high baseline functioning<sup>11,12</sup>. Further, the format is poorly suited to population-level screening and longitudinal monitoring; the requirement for in-person administration prevents seamless integration with emerging fluid biomarker pipelines, which are increasingly designed for scalable, remote deployment (e.g. AD-related plasma biomarker assays)<sup>13</sup>.

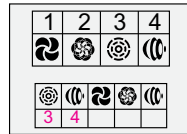
Remote, browser-based testing can circumvent these constraints while enriching psychometrics<sup>10,14,15</sup>. Digital platforms capture millisecond-level reaction-time precision and log every interaction during task execution, providing a complete behavioural trajectory that enables fine-grained

<sup>1</sup>Department of Experimental Psychology, University of Oxford, Oxford, UK. <sup>2</sup>Cognitive Disorders Clinic, John Radcliffe Hospital, Oxford, UK. <sup>3</sup>Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK. <sup>4</sup>Department of Physics, Hong Kong Baptist University, Hong Kong, China. <sup>5</sup>Neurology Residency Program, Università degli Studi di Milano, Milano, Italy. <sup>6</sup>Present address: Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK. <sup>7</sup>These authors contributed equally: Sijia Zhao, Sofia Toniolo. ✉ e-mail: [sijia.zhao@psy.ox.ac.uk](mailto:sijia.zhao@psy.ox.ac.uk)

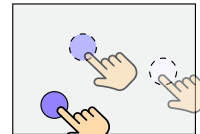
### Trail-Making Task (TMT)



### Digit-Symbol Substitution Task (DSST)      Sequence Recall Task (CORSI)

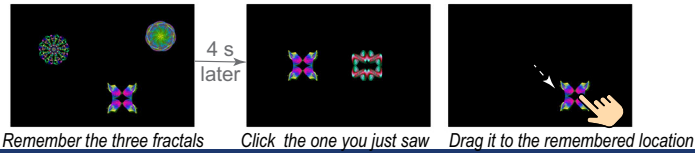


Match as many as possible in 2 min

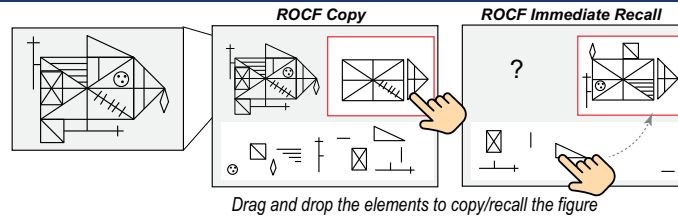


Tap the remembered locations in order

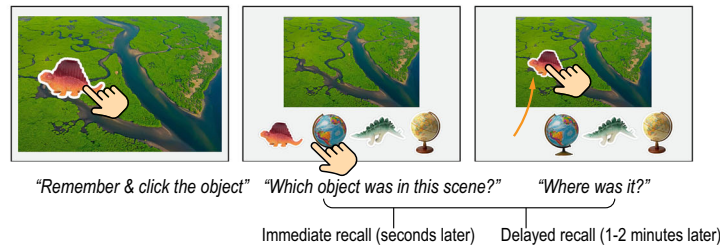
### Oxford Memory Task (OMT)



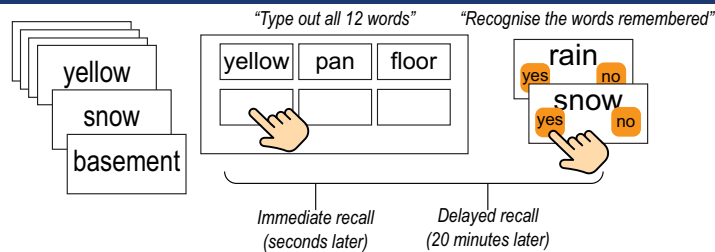
### Rey-Osterrieth Complex Figure (ROCF)



### Object-in-Scene Memory Task (OIS)



### Verbal Wordlist Recall (ALF)



cognitive stratification and supports the high-frequency sampling required to detect within-person change<sup>11</sup>.

Here, we introduce the Oxford Cognitive Testing Portal (OCTAL), a fully remote, browser-based cognitive assessment platform optimised for personal devices (Fig. 1). OCTAL comprises a suite of tasks measuring memory, attention, executive function, and visuospatial processing;

optional secondary modules can be appended to address specific research questions<sup>16,17</sup>.

To establish the validity of OCTAL across different populations and contexts, we conducted four complementary studies (see Table 1 for demographic). These studies were sequenced to progress from basic feasibility (cross-cultural comparability) to normative characterisation (lifespan

**Fig. 1 | OCTAL (Oxford Cognitive Testing Portal) task battery and outcome measures.** Note, the version reported in this manuscript was for laptops/computers or tablets. We now have optimised task for a smartphone version called mini-OCTAL, which features the same task design, but is shorter and more suitable for portrait layout. The validation of the smartphone version is still ongoing and will be reported in a separate work. (1) **Trail-Making Task (TMT)** — Participants tap onscreen nodes as rapidly as possible. A two-dot control ('Connect-2-Dots') gauges basic psychomotor speed. They then complete two 25-node numeric trails (TMT-A, a measure of attention and processing speed) and two alternating number-letter trails (TMT-B, a measure of the ability to switch between stimuli categories). Metrics are mean completion time for each condition (TMT\_Connect2Dots, TMT\_A, TMT\_B) and the TMT\_B/A ratio, a marker of executive function (cognitive flexibility). (2) **Digit-Symbol Substitution Task (DSST)** — Guided by a reference key, participants match as many symbols with correct digits as possible in 2 min. Outcomes are mean reaction time for correct substitutions (DSST\_RT) and proportion correct (DSST\_Accuracy). This tests attention and processing speed. (3) **Sequence Recall Task (CORSI)** — After observing a sequence of one, two or three highlighted circles, participants reproduce the order. Performance is summarised by mean localisation error (CORSI\_LocErr) and between-trial coefficient of variation (CORSI\_CoV). This tests spatial short-term memory. (4) **Oxford Memory Task (OMT)** — Participants view either one (easy) or three (hard) fractals and, after 4 s, drag the target fractal to its original location. Metrics are recognition accuracy (OMT\_Accuracy) and spatial displacement error (OMT\_LocError). This tests object recognition and spatial short-term memory. (5) **Rey-Osterrieth Complex Figure (ROCF)** — Participants drag-and-drop 13 predefined elements to recreate a figure ("copy") and then reproduce it from memory immediately afterwards ("recall"). Element placement is evaluated with pixel-level precision relative to a fixed central anchor, yielding proportional scores for copy (ROCF\_CopyScore

which is a measure of visuospatial perception and visuoconstructive skills) and recall (ROCF\_RecallScore which is a measure of visuospatial episodic memory) on a continuous 0–100% scale; this extends the conventional 0–36 discrete system by incorporating spatial accuracy. Memory retention, adjusted for individual visuo-motor or perceptual demands, is expressed as the ratio  $ROCF\_Remember = ROCF\_RecallScore \div ROCF\_CopyScore$  in percentage. See Methods, 'Ethical and intellectual property considerations', for stimulus rationale and copyright details. (6) **Object-in-Scene Memory Task (OIS)** — participants identify an object placed on a random location on a natural scene and subsequently choose the object and report its location after a short (seconds) and longer (1–2 min) delay. Metrics are recognition accuracy and localisation error for short-term memory (OIS\_STM\_Accuracy, OIS\_STM\_LocErr) and long-term memory (OIS\_LTM\_Accuracy, OIS\_LTM\_LocErr). (7) **Verbal Wordlist Recall (ALF)** — Participants study a list of 12 words, then complete immediate free recall (nWords\_immediate) and recognition (RecAcc\_immediate). After ~30 min filled with intervening tasks, they complete delayed free recall (nWords\_delayed) and recognition (RecAcc\_delayed). This task indexes verbal episodic memory. *In OCTAL documentation, this task has sometimes been referred to as Accelerated Long-Term Forgetting (ALF), as it was adapted from ALF paradigms. However, the present implementation uses a short-delay adaptation (<1 h) that differs from conventional ALF protocols, which typically employ ≥24 h delays. We retained the abbreviation ALF to ensure consistency across OCTAL publications and documentation.* All elements of this figure were created by the authors in Adobe Illustrator to illustrate the OCTAL task design. Visual elements are either original to the OCTAL platform, free for academic use (objects from Brady et al.<sup>47</sup>), or used under paid licence (e.g., icons and photos via The Noun Project). The overall task design and figure composition are the intellectual work of the authors. See 'Ethical and intellectual property considerations' for further details.

**Table 1 | Demographic characteristics and ACE-III and OCTAL performance across the four studies**

Study	Group	N	Gender	Age	Education	ACE-III total score	OCTAL ROCF RecallScore (%)
1	English	194	F87 M105 [N = 192]	31.6 (6.9) [N = 194]	15.0 (2.0) [N = 93]	N.A.	92.4 (10.6) [N = 192]
	Chinese	167	F96 M71 [N = 167]	29.9 (5.7) [N = 167]	16.5 (1.5) [N = 167]	N.A.	95.1 (7.7) [N = 166]
2	HC	1109	F626 M479 [N = 1106]	54.1 (12.7) [N = 1109]	15.3 (2.1) [N = 374]	N.A.	86.6 (13.7) [N = 1102]
3	HC	97	F53 M44 [N = 97]	60.7 (17.6) [N = 97]	17.9 (4.0) [N = 93]	97.1 (2.7) [N = 97]	86.5 (13.8) [N = 94]
	SCD	32	F12 M17 [N = 29]	61.4 (8.5) [N = 31]	15.9 (3.6) [N = 30]	95.4 (5.7) [N = 32]	88.5 (17.9) [N = 27]
	MCI	11	F4 M6 [N = 10]	67.0 (6.5) [N = 11]	12.6 (3.2) [N = 11]	88.5 (5.4) [N = 11]	72.3 (19.5) [N = 8]
	AD	54	F29 M24 [N = 53]	69.9 (8.9) [N = 51]	14.8 (4.6) [N = 47]	70.6 (16.7) [N = 54]	46.4 (28.0) [N = 43]
	Total	194	F98 M91 [N = 189]	63.7 (14.4) [N = 190]	16.5 (4.4) [N = 181]	88.9 (14.9) [N = 194]	76.1 (25.8) [N = 172]
4	HC	59	F33 M26 [N = 59]	66.8 (11.6) [N = 59]	16.2 (3.2) [N = 56]	97.4 (1.9) [N = 58]	89.0 (11.6) [N = 59]
	SCD	23	F9 M13 [N = 22]	62.0 (9.0) [N = 23]	15.7 (3.4) [N = 22]	94.0 (4.2) [N = 5]	93.9 (7.8) [N = 21]
	MCI	5	F2 M3 [N = 5]	67.9 (4.7) [N = 5]	10.4 (0.5) [N = 5]	90.2 (4.5) [N = 2]	76.2 (17.8) [N = 5]
	AD	31	F18 M12 [N = 30]	70.1 (9.0) [N = 31]	14.3 (3.9) [N = 27]	72.2 (12.6) [N = 24]	43.1 (23.6) [N = 28]
	Total	118	F62 M54 [N = 116]	66.8 (10.5) [N = 118]	15.3 (3.6) [N = 110]	90.3 (13.0) [N = 89]	78.0 (25.3) [N = 113]

For the continuous variables (age, number of years education etc) data are presented as mean (1 standard deviation). Square brackets denote the number of participants with valid data (i.e., excluding the missing data). N.A. indicates data not applicable. For Study 4, ACE-III and OCTAL scores are averaged across all assessments. Further details are provided in the manuscript and Methods.

data), to clinical sensitivity, and finally to longitudinal stability. Specifically: (1) Study 1: Cross-cultural generalisability (N = 361). We compared English- and Chinese-speaking young adults (18–45 years) in the United Kingdom and China to test whether performance is comparable across languages and cultural contexts. (2) Study 2: Lifespan scalability (N = 1109). We assessed age-related change across cognitive domains in a large UK sample of adults aged 45–85 years. (3) Study 3: Clinical utility (N = 194). We evaluated whether OCTAL detects cognitive impairment defined by ACE-III and whether it distinguishes Alzheimer's disease or mild cognitive impairment (MCI) from subjective cognitive decline (SCD). (4) Study 4: Test–retest reliability (N = 118). We examined performance stability across repeated assessments over 6 months in both healthy and clinical participants.

Together, these studies were designed to evaluate OCTAL's feasibility, scalability, sensitivity and robustness as a practical tool for digital screening

and longitudinal monitoring, suitable for both clinical and research applications. Importantly, OCTAL is designed with researchers in mind. It produces data-rich outputs that are readily compatible with a wide range of digital platforms, including survey tools (e.g. Qualtrics), online participant recruitment systems (e.g. Prolific, MTurk), experimental testing platforms (e.g. Pavlovia, Gorilla), and complementary cognitive tools (e.g. Sea Hero Quest<sup>18</sup>, Cognitron<sup>19</sup>). OCTAL supports both remote and in-person deployment and enables high-resolution cognitive phenotyping across diverse populations and study designs.

## Results

### Study 1: International scalability in young adults

The goal of Study 1 was to test whether OCTAL provides comparable performance across different language and cultural groups, thereby

**Table 2 | Demographic and cognitive performance comparisons between English- and Chinese-speaking participants (Study 1)**

Metric	Mean (SD)		Bootstrap <i>p</i> -value	
	English	Chinese	Below - 1 SD	Above 1 SD
<i>N</i>	194	167		
Gender (% Female)	44.9	57.5		
Age (years)	31.60 (6.88)	29.88 (5.72)	N.S.	N.S.
Education (years)	14.96 (2.02)	16.46 (1.52)	N.S.	N.S.
ROCF_CopyScore	99.07 (4.75)	99.26 (2.36)	N.S.	N.S.
ROCF_CopyDuration	105.87 (38.54)	117.85 (52.62)	N.S.	N.S.
ROCF_RecallScore	92.36 (10.60)	95.07 (7.73)	N.S.	N.S.
ROCF_Remember	93.15 (9.38)	95.77 (7.36)	N.S.	N.S.
ROCF_RecallDuration	88.92 (39.91)	100.97 (56.53)	N.S.	N.S.
OMT_Acc_Easy	96.95 (5.85)	98.86 (2.46)	N.S.	N.S.
OMT_Acc_Hard	85.88 (10.16)	89.96 (8.76)	N.S.	N.S.
OMT_LocErr_Easy	0.08 (0.05)	0.07 (0.03)	N.S.	N.S.
OMT_LocErr_Hard	0.16 (0.07)	0.15 (0.07)	N.S.	N.S.
OMT_Acc	91.42 (6.41)	94.41 (4.84)	N.S.	N.S.
OMT_IdeRT	1.73 (0.49)	2.28 (0.79)	N.S.	N.S.
OMT_LocErr	0.12 (0.05)	0.11 (0.04)	N.S.	N.S.
OMT_LocRT	2.10 (0.77)	2.50 (1.86)	N.S.	N.S.
OMT_TargetDetection	72.74 (13.84)	82.12 (12.55)	N.S.	N.S.
OMT_Misbinding	24.90 (10.91)	17.92 (11.88)	N.S.	N.S.
OMT_Guessing	14.73 (8.98)	8.88 (7.16)	N.S.	N.S.
OIS_STM_SemanticAcc	99.65 (2.86)	99.88 (0.77)	N.S.	N.S.
OIS_STM_Acc	97.33 (4.76)	97.49 (7.68)	N.S.	N.S.
OIS_STM_IdeRT	1.73 (0.24)	2.57 (0.92)	N.S.	<i>p</i> < 0.001
OIS_STM_LocErr	0.21 (0.05)	0.18 (0.05)	N.S.	N.S.
OIS_STM_LocRT	5.71 (0.39)	6.15 (3.64)	N.S.	N.S.
OIS_LTM_SemanticAcc	86.11 (16.75)	87.10 (15.28)	N.S.	N.S.
OIS_LTM_Acc	77.94 (19.79)	81.29 (18.42)	N.S.	N.S.
OIS_LTM_IdeRT	6.65 (1.90)	6.27 (1.82)	N.S.	N.S.
OIS_LTM_LocErr	0.21 (0.05)	0.20 (0.06)	N.S.	N.S.
OIS_LTM_LocRT	11.18 (1.12)	10.33 (3.49)	N.S.	N.S.
DSST_nCorrectResponse	52.11 (14.08)	51.88 (11.02)	N.S.	N.S.
DSST_RT	2.37 (2.27)	2.20 (0.47)	N.S.	N.S.
DSST_pHit	97.91 (3.57)	98.95 (1.90)	N.S.	N.S.
TMT_Connect2	3.72 (0.98)	2.56 (0.56)	N.S.	N.S.
TMT_A	20.28 (9.24)	20.78 (7.10)	N.S.	N.S.
TMT_B	29.01 (11.54)	31.53 (9.09)	N.S.	N.S.
TMT_BdA	147.65 (35.30)	156.64 (40.80)	N.S.	N.S.

Bootstrap-based practical significance was evaluated with a permutation procedure (10,000 iterations). For each metric, lower and upper thresholds were defined as 1 SD below and 1 SD above the mean of the English-speaking cohort, respectively. To determine whether the Chinese group (*N* = 167) differed meaningfully from these thresholds, *N* observations were resampled with replacement from the Chinese data, the mean was calculated, and this process was repeated 10,000 times. A difference was deemed significant when ≥ 97.5% of resampled means lay below the lower threshold or above the upper threshold (*p* < 0.025). Non-parametric group comparison for sample mean (Mann–Whitney *U* tests) are reported in Supplementary Table 1. Reported values include the *U* statistics, the effect size (*f*) and Bonferroni-corrected *p* values.

evaluating its cross-cultural generalisability. To address this, we recruited 361 young adults (aged 18–44), including 194 native English speakers who completed the English version and 167 native Chinese speakers who completed the Chinese version of OCTAL. Note that the Chinese-language OCTAL battery currently excluded the Verbal Memory Word-List Recall task because text-entry methods are not standardised across age groups in China: younger adults typically use pinyin keyboards, whereas many older adults rely on stroke-based input (e.g. Wubi), handwriting interfaces, or have limited typing skills, preventing equitable administration.

All participants completed OCTAL remotely on their own laptops or desktop computers, regardless of operating system or browser. The platform was developed in PsychoJS, facilitating straightforward modification by researchers with minimal coding experience, and hosted on Pavlovia. Recruitment was conducted via Prolific (English version with UK residents) and Credamo (Chinese version with residents of mainland China). Table 2 summarises demographic characteristics; no meaningful differences emerged between the two cohorts (except Chinese participants were slightly more educated).

Cognitive performance on core OCTAL metrics was largely comparable between the English-speaking and Chinese-speaking groups (Table 2). Indeed, for the majority of key metrics, the mean performance of the Chinese cohort fell within the normative range established by the English-speaking cohort (defined as  $\pm 1$  standard deviation of the English group's mean, determined via a bootstrap permutation procedure; see Table 2).

However, when examining reaction times, a significant group difference that exceeded this normative threshold was observed for the identification response time in the immediate recall phase of the Object-in-Scene (OIS) task (OIS\_STM\_IdeRT). Chinese participants took an average of 2.57 s (SD = 0.92) for this measure, compared to 1.73 s (SD = 0.24) for English-speaking participants. This difference meant the Chinese group's mean reaction time was significantly greater than one standard deviation above the English group's mean. A speed-accuracy trade-off was an unlikely explanation for this specific finding, as the Chinese group showed only a marginal 0.16% advantage in OIS identification accuracy (97.49% vs. 97.33%), a difference too small to account for the substantial variance in response time. Meanwhile, although insignificant, several reaction time measures indicated slower responses in the Chinese group; the reason behind this systematically slower response is not immediately clear. We will come back to this in 'Discussion'.

A noteworthy difference emerged in Trail Making Test Part B (TMT-B), where Chinese participants were slower than the UK cohort (China:  $M = 31.5$  s,  $SD = 9.1$ ; UK:  $M = 29.0$  s,  $SD = 11.4$ ;  $U = 29,565$ ,  $p = 0.01$ ,  $r = 0.20$ ), despite having learned English since childhood. In contrast, TMT-A performance was comparable between the groups ( $U = 31,307$ ,  $p = 1.00$ ,  $r = 0.10$ ). A similar pattern was observed in Study 2, using the combined dataset of 1109 UK residents across the lifespan: non-native English speakers ( $n = 191$ ) demonstrated significantly greater TMT-B latency from age 55 onwards relative to native speakers ( $n = 918$ ), an effect that persisted after controlling for TMT-A performance (see Supplementary Fig. 1).

Could the variance in OCTAL performance be merely explained by difference in devices? During development of the platform we considered whether display size might confound OCTAL scores. The Rey-Osterrieth Complex Figure (ROCF; see 'Methods', 'Ethical and intellectual property considerations', for stimulus rationale and copyright details) was the prime candidate for such an effect: on a large screen, repeatedly dragging elements could prolong completion time, whereas on a small screen limited motor precision might hamper accurate placement. To isolate memory from such visuomotor factors we required participants to copy the figure first and then reproduce it from memory. Performance is therefore indexed both by the raw Copy score (i.e. a 100% scale showing how precise are ROCF's elements located relative to each other) and by the Recall/Copy ratio, which normalises for any screen-related copying differences. To test this hypothesis, we expanded the mainland-Chinese cohort with 43 additional adults aged 45–70, yielding  $N = 210$ . All participants used either a laptop or a desktop computer; tablets were disallowed. Although device type and pointing method were not logged explicitly, the browser captured screen width and height (pixels) at test.

Supplementary Fig. 2 ranks the correlations between screen width and all 34 OCTAL metrics. Only ROCF Copy was affected: wider screens were associated with longer completion times ( $\rho = -0.26$ ,  $p = 0.01$ ) and higher precision ( $\rho = 0.26$ ,  $p = 0.01$ ). Time and accuracy were, however, independent ( $\rho = -0.04$ ,  $p = 0.61$ ), discounting a simple speed-accuracy trade-off. Supplementary Fig. 3 shows analogous results for screen height. Copy score again correlated with screen height ( $\rho = 0.30$ ,  $p = 0.0004$ ), and the strength of the width and height associations did not differ (Fisher's  $z = -1.60$ ,  $p = 0.11$ ). Time taken to complete ROCF copy, by contrast, was unrelated to height.

Crucially, how well ROCF was recalled (ROCF Recall score) did not correlate with screen dimensions (width  $\rho = 0.13$ ; height  $\rho = 0.11$ ; both  $p = 1$ ). The memory index (Recall/Copy  $\times 100\%$ ) was likewise insensitive to screen size (width  $\rho = 0.06$ ; height  $\rho = 0.02$ ; both  $p = 1$ ). Thus, while larger displays may marginally facilitate more precise copying—every participant nevertheless scored above 90%—they do not influence the amount of

information retained. OCTAL memory measures are therefore robust to the variability in laptop and desktop screen dimensions encountered in remote testing.

## Study 2: Lifespan scalability and sensitivity to age-related differences

Study 1 demonstrated that culturally diverse young adults can complete OCTAL unaided, supporting its feasibility for remote administration in participants below 45 years of age. Building on this, Study 2 tested whether adults in mid- to late-life can also complete the platform independently and whether OCTAL captures the expected cross-sectional associations between age and cognitive performance across domains.

We recruited 1109 UK residents via Prolific, of whom 905 were aged 45–89 years. All assessments were completed remotely without supervision, and only a small proportion of participants failed the attention checks (see 'Methods'), indicating that OCTAL is broadly feasible for unsupervised administration across the adult lifespan. Z-scores for all metrics were computed relative to the 18–39-year reference group from Study 1, unless otherwise stated. This design allowed us to evaluate both the feasibility of unsupervised testing and the sensitivity of OCTAL to age-related variation in memory, attention, executive function, and visuospatial processing.

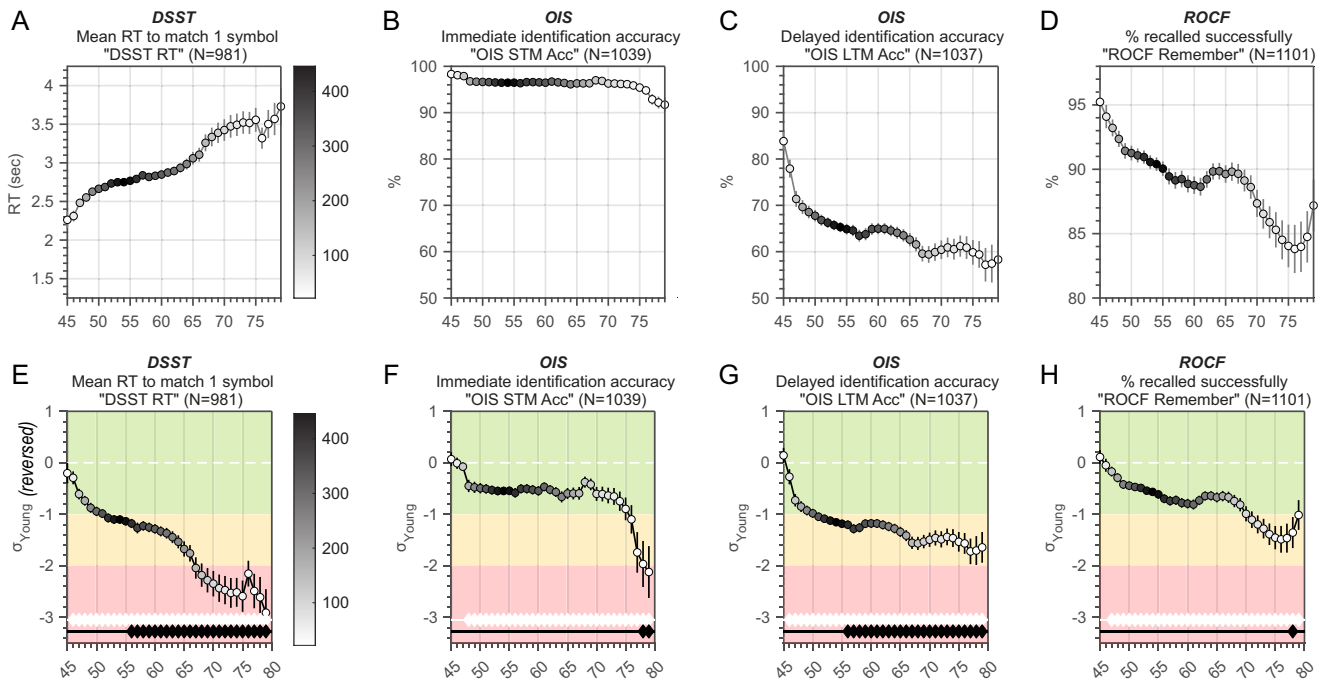
Analysis of cognitive performance revealed that most metrics correlated significantly with age (Supplementary Table 4), with the strongest associations observed for executive functions. As shown in Fig. 2, the Digit Symbol Substitution Test (DSST) reaction time increased steadily with age, diverging from the young-adult reference band by the mid-forties and crossing 1 SD below the reference mean by the mid-fifties (Fig. 2A, E). Error rates, however, remained low until after age 65 (Supplementary Fig. 5A, B). Trail Making Test (TMT-A and TMT-B) completion times also slowed progressively, exceeding the  $-1$  SD boundary in the mid-to-late fifties (Supplementary Fig. 5E–H). Performance on the visuomotor control task showed only a modest latency increase (Supplementary Fig. 5C) and remained within normal limits across the lifespan (Supplementary Fig. 5D).

For memory, immediate identification accuracy ('which object was in this photo?' in the Object-in-Scene task) remained close to ceiling across much of adulthood and within the normative range until the late seventies (Fig. 2B, F). By contrast, when memory for the same objects was tested after a brief delay, accuracy declined sharply from the mid-fifties onwards (Fig. 2C, G). Performance on the Oxford Memory Test (OMT, Supplementary Fig. 5I, J) and delayed recall of the Rey-Osterrieth Complex Figure (Fig. 2D, H) also remained within the normative range until the late seventies.

To compare the relative rates of change across domains, we fitted straight lines to the Gaussian-smoothed, age-weighted means of each metric and ranked the resulting slopes (Fig. 3). The steepest decline was observed for DSST reaction time ( $-0.073$  z-units year $^{-1}$ ,  $R^2 = 0.95$ ), closely followed by the ROCF copy score ( $-0.070$ ). Slopes for ROCF immediate recall ( $-0.035$ ) and OIS delayed identification ( $-0.031$ ) were roughly half as steep. Full slope estimates and  $R^2$  values are provided in Supplementary Table 5, with individual fits shown in Supplementary Fig. 6.

Verbal memory showed minimal age-related differences. In the Verbal Wordlist Recall (abbreviated ALF), immediate recall declined by only  $-0.009$  z-units year $^{-1}$  and delayed recall by  $-0.020$  z-units year $^{-1}$ . Each list contained one word from twelve semantic categories, pseudo-randomised across 100 sets. Participants studied a list until they achieved at least 50% recall accuracy ( $\geq 6$  words), ensuring adequate encoding for delayed recall. Most unimpaired participants reached this criterion on the first attempt, and the number of learning trials was logged but not included in subsequent analyses (see 'Methods' for full details). Adults aged 45+ recalled on average slightly more words ( $\sim 9$  items) than younger participants, and despite smaller samples in the late sixties and seventies, most scores remained within the normative range and did not differ significantly from the  $-1$  SD threshold (Fig. 3).

A key finding from this large healthy cohort ( $N = 1109$ ) was the surprisingly modest strength of the correlations between most cognitive metrics, with the majority of Spearman's  $\rho$  values falling below 0.4 (Fig. 4, or



**Fig. 2 | Cross-sectional associations between age and OCTAL metrics in a healthy population (Study 2, N = 1109).** The top panel displays raw scores for the DSST (A), OIS Short-Term (B) and Long-Term (C) recall, and ROCF memory (D) across ages 45–80 years. Curves are smoothed with a Gaussian kernel (bandwidth = 2.5 years). Marker colour encodes the number of observations contributing to each point, as shown by the colour bar on the right. The bottom panel (E–H) presents the same metrics normalised to the young-adult reference group (18–39 years). The scale demarcates the normative range ( $\pm 1$  SD, green zone), moderate deviation

(1–2 SD, orange zone), and marked impairment ( $\geq 2$  SD, red zone). Reaction-time variables—such as DSST mean RT in A—are sign-reversed so that increasingly negative z-scores indicate greater executive decline. White diamonds along the horizontal white guideline mark age groups that differ significantly from the young-adult mean, whereas black diamonds indicate significant deviation beyond  $-1$  SD (i.e., the boundary between the green and orange zones). See Supplementary Fig. 5 for additional metrics.

see the full correlation matrix with unadjusted  $p$  values in Supplementary Fig. 7).

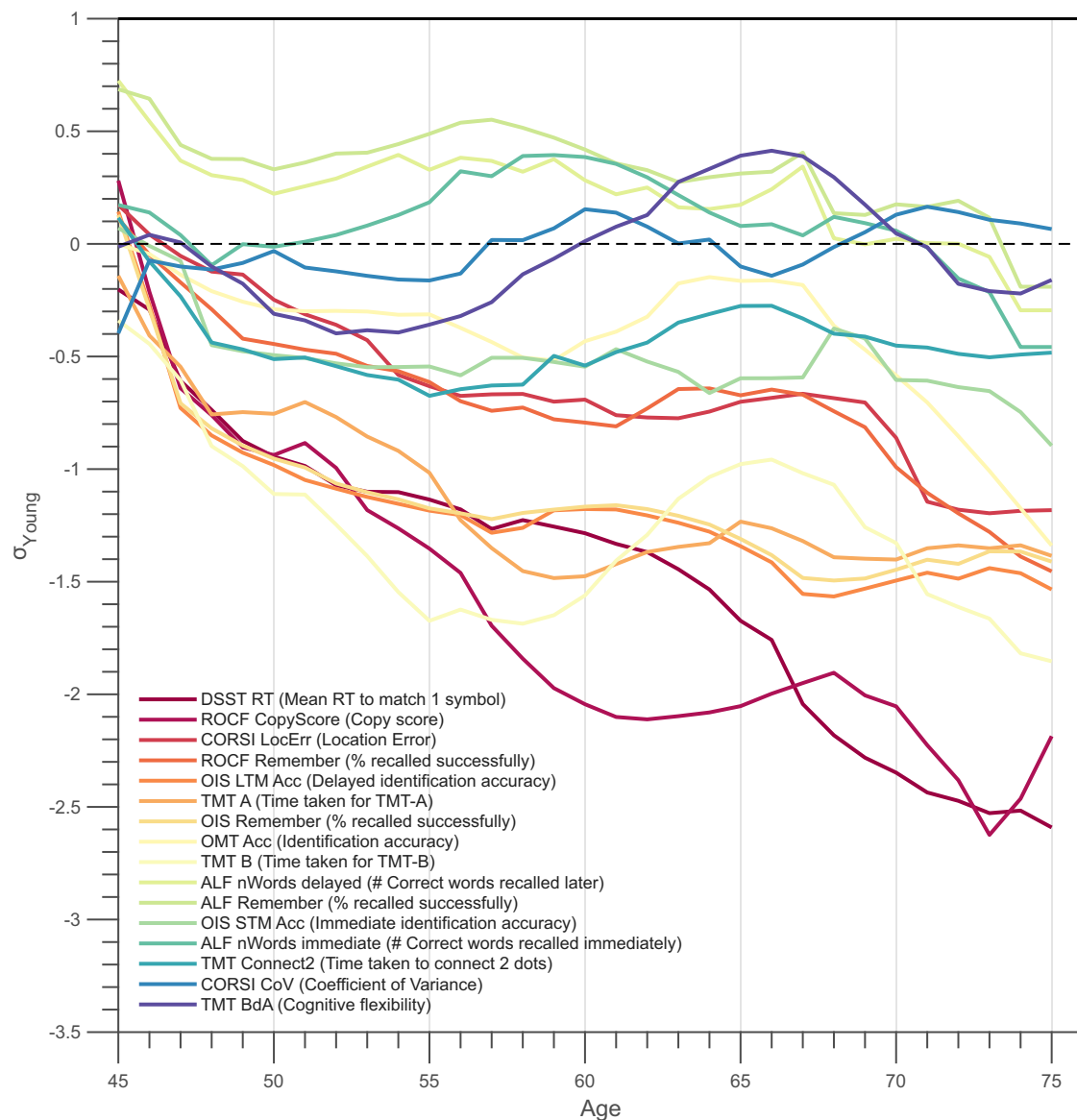
To understand the structure of these relationships at different levels of granularity, we employed two complementary analytical approaches. First, a high-level exploratory factor analysis (EFA) was used to identify the main organising principles of the cognitive data. This analysis revealed a clear and interpretable two-factor structure (using varimax rotation) corresponding to the classical domains of (1) Memory and (2) Executive Function (Supplementary Fig. 4). The analysis confirmed the suitability of the data for factorisation (KMO = 0.68; Bartlett’s Test:  $\chi^2 = 539.14, p < 0.001$ ) and the sufficiency of a two-factor solution ( $\chi^2 = 235.54, df = 134, p < 0.001$ ). To allow for inter-factor correlations, we repeated the analysis using promax rotation, which yielded a similar 2-factor structure. Memory and executive function were weakly negatively correlated (factor correlation =  $-0.2$ ), possibly reflecting individual differences in cognitive strategies that prioritise either accuracy or speed, reinforcing their distinctness.

Second, to investigate the finer-grained relationships within and between these broad domains, we constructed a weighted, undirected network from 25 key metrics and applied a single-pass Louvain-style modularity maximisation algorithm (Fig. 5, see ‘Methods’ for details). This method does not require strong correlations but identifies communities of nodes that are more densely interconnected with each other than with the rest of the network. The analysis partitioned the network into six distinct modules. These modules appear to reflect both task-specific effects and broader cognitive functions:

1. Executive function: This module’s central node was TMT-A, and it included other members such as DSST\_RT, DSST Accuracy, and TMT-Connect2Dots.
2. Completion time: Centred on the ROCF copy duration, this two-node module also contained the ROCF recall duration, likely reflecting task strategy and motor speed.

3. CORSI module: This module was highly task-specific, consisting only of the CORSI localisation error (the central node) and its coefficient of variation.
4. Short-term memory: This module contained all metrics from the OMT, with OMT\_Guessing as the central node.
5. Visuospatial precision: With OIS long-term memory localisation error as its central node, this module primarily grouped metrics of spatial accuracy, including the ROCF Copy Score and OIS short-term memory localisation error. Interestingly, it also contained the TMT B/A ratio, a measure of cognitive flexibility, suggesting a link between executive control and the performance of precise visuospatial tasks.
6. Episodic memory: This was a large, cross-modal memory module with delayed verbal free recall (ALF\_nWords\_delayed) as its central node. It connected the other verbal memory (ALF) metrics with visuospatial memory retention (ROCF\_Remember) and both short- and long-term object identification accuracy (OIS\_ST-M\_Acc and OIS\_LTM\_Acc).

To confirm the stability and robustness of the detected community structure, we ran the Louvain algorithm for 10,000 iterations. The 6-module solution demonstrated remarkable robustness, being obtained in 83.99% of these independent runs. This high consistency, despite the algorithm’s internal randomisation (due to arbitrary node ordering), indicates that the 6-module partition is a highly stable and recurring local maximum in the modularity landscape of our cognitive network. The modularity  $Q$  for this dominant solution was consistently low ( $Q = 0.13 \pm 0.00$ ), suggesting that while clear communities exist, the inter-module connections are not vastly sparser than expected by chance for this specific network, or that the algorithm tends towards a coarse-grained partitioning given the correlation strengths.



**Fig. 3 | Comparative slopes of cross-sectional age-performance associations across OCTAL metrics (Study 2).** Lines correspond to the fitted curves from the bottom panels of Fig. 2, here combined into a single plot for direct comparison across metrics. Curves are colour-coded and ordered by the magnitude of their negative age slope: warmer (redder) lines denote steeper decline from 45 to 75 years (larger annual reduction in z-score), whereas cooler (bluer) lines denote shallower decline.

The y-axis shows performance expressed in z-scores relative to the 18–39 year reference group from Study 1. See Supplementary Table 5 for slope estimates and  $R^2$  values of the linear fits (see ‘Methods’ for details). Note that ALF refers to the Verbal Wordlist Recall task, implemented here as a short-delay adaptation (<1 h) of traditional ALF paradigms.

Taken together, these analyses resolve the apparent paradox of finding structured modules despite weak inter-correlations. The EFA reveals two broad, classical cognitive domains, while the network analysis demonstrates that these domains are not monolithic. Instead, they are composed of multiple, weakly correlated sub-components. This finding is a key strength of the OCTAL battery. It indicates that the individual tasks provide high-resolution, non-redundant information, enabling fine-grained cognitive phenotyping.

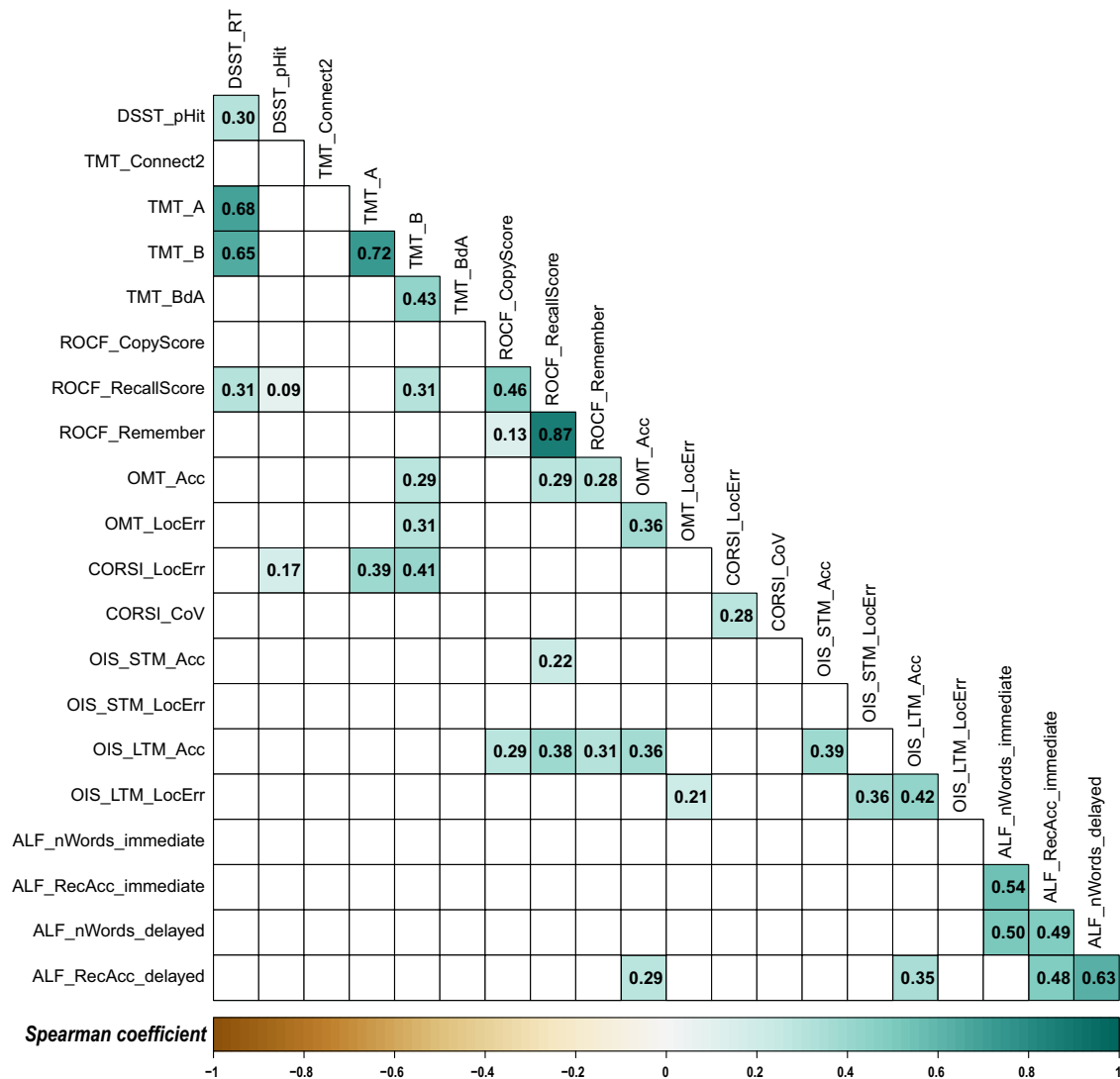
**Study 3: Does remote testing performance align with in-person standard cognitive assessment?**

Having demonstrated in Study 1 that OCTAL is feasible across languages and in Study 2 that it captures expected age-related differences, the next step was to test its clinical validity. We therefore evaluated whether OCTAL performance aligns with the ACE-III, a widely used in-person screening tool, and whether brief OCTAL composites can match or surpass ACE-III in

discriminating between SCD, MCI, and Alzheimer’s disease dementia. This design was intended to establish concurrent validity against a clinical benchmark while evaluating potential advantages in brevity and domain specificity.

A total of 194 individuals —97 cognitively healthy controls, 54 with early Alzheimer’s disease dementia, 11 with MCI and 32 with SCD— completed OCTAL remotely and the ACE-III during an in-person visit ( $\approx 15\text{--}20$  min administration), with test order pseudo-balanced across the cohort.

Spearman rank correlations revealed a strong association between OCTAL metrics and the ACE-III total score (Table 3). The metrics from the wordlist recall task (coded as ALF) showed the highest concordance ( $\rho = 0.61\text{--}0.74$ ). After Bonferroni correction, three variables did not reach significance ( $\alpha = 0.05$ ): Digit-Symbol hit rate ( $\rho = 0.15$ ), CORSI Coefficient of variance in localisation accuracy (CORSI\_CoV,  $\rho = 0.06$ ) and Object-in-Scene short-term localisation accuracy (OIS\_STM\_LocErr,  $\rho = 0.13$ ).



**Fig. 4 | Metric-by-metric correlations among OCTAL indices in healthy adults (Study 2).** The matrix displays Spearman coefficients ( $\rho$ ) for all pairwise comparisons; only correlations that remain significant after Bonferroni correction ( $p < 0.05$ ) are shown. All metrics are directionally aligned so that lower scores indicate greater

cognitive impairment. Square colour, mapped to the accompanying colour bar, encodes the magnitude and sign of  $\rho$ , and each cell is labelled with its exact value. Full correlation coefficients and unadjusted  $p$  values are reported in Supplementary Fig. 6.

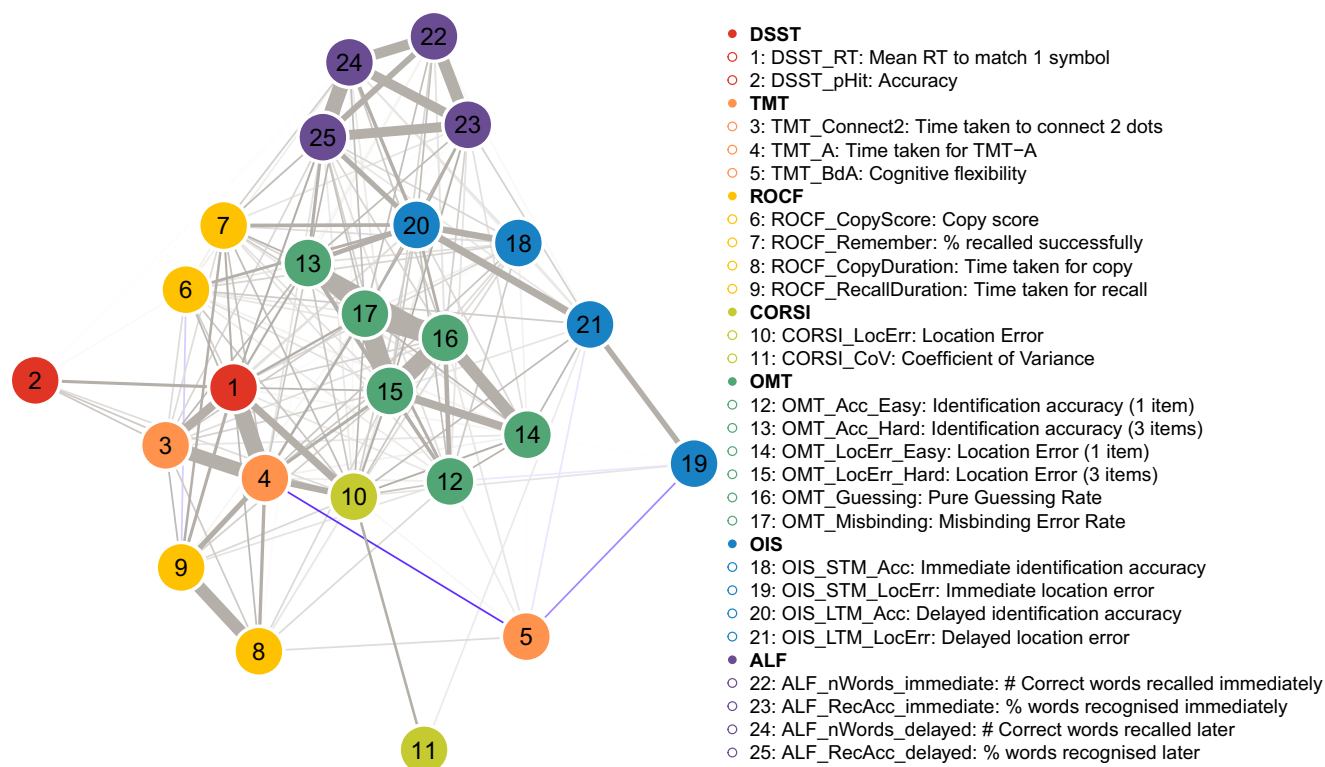
Because ACE-III subscores are highly inter-correlated (mean inter-correlation  $\rho = 0.74$ ; Supplementary Table 6), partial correlations were computed while controlling for the remaining four domains (Fig. 6). Trail-Making Test part A, Trail-Making cognitive flexibility (ratio between time taken to complete trails B and A) and object identification accuracy in immediate recall of OIS memory task (short-term memory) were uniquely associated with the ACE-III Attention subscore, whereas the immediate and delayed wordlist recall mapped specifically to ACE-III Memory. Identification accuracy in OMT task and both the short- and long-term phases of OIS showed moderate, selective correlations with ACE-III Fluency. No OCTAL metric exhibited a domain-specific relation with ACE-III Visuospatial or Language.

A machine learning-based feature-ranking algorithm identified ROCF immediate recall score, OIS object identification accuracy in delayed recall (long term memory) and time taken to complete TMT-A as the three most informative predictors of ACE-III-defined cognitive impairment (cutoff = 88, Fig. 7A). Each of these top metrics achieved excellent discrimination in isolation, with area-under-the-curve (AUC) values  $\geq 0.88$  (Fig. 7B) and optimal raw-score thresholds of: ROCF immediate recall score  $< 68.9\%$  (AUC = 0.91, sensitivity = 79.1%, specificity = 89.9%), OIS object

identification accuracy  $< 45\%$  (AUC = 0.90, sensitivity = 90.0%, specificity = 83.3%) and Trail-Making Test part A  $> 32.8$  s (AUC = 0.88, sensitivity = 85.0%, specificity = 84.9%). Performance was equivalent when z-scores were substituted for raw scores (ROCF recall score:  $< -1.0$  SD, AUC = 0.89, sensitivity = 81.4%, specificity = 86.0%; OIS identification accuracy:  $< -0.8$  SD (AUC = 0.89, sensitivity = 87.5% and specificity = 83.3%; TMT A  $< -0.9$  SD (AUC = 0.88, sensitivity = 78.9% and specificity = 84.7%). Although not ranked as top, the 2-min executive function task—the Digit-Symbol Substitution Test—alone also produced strong classification (AUC = 0.84; Fig. 7B).

Combining these top three tasks (ROCF, OIS & TMT), which in total take about 20 min of testing, yielded an AUC of 0.96, with sensitivity = 87.1% and specificity = 96.2% (Fig. 7B). A rapid 5-min set (ROCF + DSST) also performed very well (AUC = 0.92, sensitivity = 81.5%, specificity = 87.9%).

To test robustness to diagnostic thresholding, analyses were repeated at ACE-III  $< 85$  and  $< 82$ <sup>20</sup>. Discrimination remained high and the same metrics consistently ranked top (Supplementary Fig. 8). For the 20-min set, cross-validated LASSO AUCs were 0.93 (95% CI 0.91–0.95) at  $< 88$ , 0.95 (0.92–0.96) at  $< 85$ , and 0.94 (0.90–0.96) at  $< 82$ ; Random Forest (RF) out-of-



**Fig. 5 | Network plot of relationships among OCTAL key metrics in the healthy population (Study 2).** Each node represents one metric and is colour-coded by task (legend ordered by task). Edge thickness and inter-node distance scale with the absolute Spearman correlation coefficient ( $\rho$ ): shorter, thicker edges denote stronger associations. Edges are displayed only when  $|\rho| \geq 0.10$ . All metrics are directionally aligned so that lower scores signify greater cognitive impairment. Grey edges indicate positive correlations, where better performance on metric A accompanies better

performance on metric B. Unsurprisingly, most relationships are positive. Only two edges are negative (highlighted in violet): between metric 5 (TMT cognitive flexibility, defined as TMT-B/TMT-A) and metric 4 (TMT-A completion time), and between metric 5 and metric 19 (OIS short term location error). Full correlation coefficients and unadjusted  $p$  values are reported in Supplementary Fig. 6. The correlation matrix with multiple comparison correction applied is in Fig. 5.

bag AUCs were 0.97 (0.96–0.98), 0.98 (0.97–0.98), and 0.96 (0.95–0.97), respectively. For the 5-min set, LASSO AUCs were 0.89 (0.86–0.92), 0.92 (0.89–0.94), and 0.93 (0.90–0.95); RF AUCs were 0.90 (0.86–0.92), 0.92 (0.90–0.94), and 0.92 (0.91–0.94). Sensitivity and specificity were balanced across thresholds (e.g. 20-min at <88: LASSO 0.84/0.95; RF 0.90/0.92; 5-min at <88: LASSO 0.85/0.82; RF 0.91/0.77).

Stability analysis confirmed robustness. Within-method resampling yielded mean pairwise Spearman  $\rho$  of 0.92–0.96 for LASSO and 0.96–0.97 for RF at <88, remaining comparably high at <85 and <82. Cross-cut-off stability of feature-importance rankings was also strong (20-min set: LASSO  $\rho = 0.95$  for <88 vs <85, 0.83 for <88 vs <82; RF  $\rho = 0.90$ –1.00; 5-min set: LASSO  $\rho = 0.94$ –1.00; RF  $\rho = 0.60$ –1.00; Supplementary Table 7).

Together, these findings demonstrate that (i) OCTAL’s top predictive metrics are robust across clinically used ACE-III thresholds, and (ii) both 20-min and 5-min composites deliver generalisable discrimination of cognitive impairment.

Having established that OCTAL reliably reproduces ACE-III classifications across cut-offs, we next asked whether its metrics could also discriminate between clinical diagnostic groups. In a subset of 97 participants (65 with Alzheimer’s disease dementia or MCI; 32 with SCD), the 20-min OCTAL composite yielded an AUC of 0.97, significantly outperforming the ACE-III total score (AUC = 0.92; DeLong  $p = 0.042$ ; Fig. 7C). A compressed 5-min OCTAL composite matched the accuracy of ACE-III (AUC = 0.92 for both; no significant difference:  $p = 0.69$ ).

To further probe its clinical utility, we used multiple linear regression to determine if performance on the 20-min OCTAL battery (comprising the ROCF, OIS, and TMT) could predict scores on the in-person ACE-III. The model proved to be a strong predictor of the ACE-III total score ( $r = 0.83$ ,

$p < 0.001$ ). High predictive accuracy was also achieved for the Attention, Memory, Visuospatial, and Fluency subscores, with correlations ranging from  $r = 0.75$  to 0.80 (all  $p < 0.001$ ). A significant, though more moderate, correlation was found for the Language subscore ( $r = 0.57$ ,  $p < 0.001$ ); this may be partly explained by the high degree of inter-correlation between the ACE-III subscores themselves, where performance on the language domain is also associated with other cognitive functions. The close correspondence between OCTAL-predicted and actual clinician-administered scores for all domains is illustrated in Fig. 8.

Collectively, these results demonstrate that remote OCTAL assessment aligns closely with in-person ACE-III performance, provides domain-specific cognitive insights and accurately discriminates pathological from non-pathological ageing within brief testing times.

#### Study 4: Six-month test–retest reliability across three visits

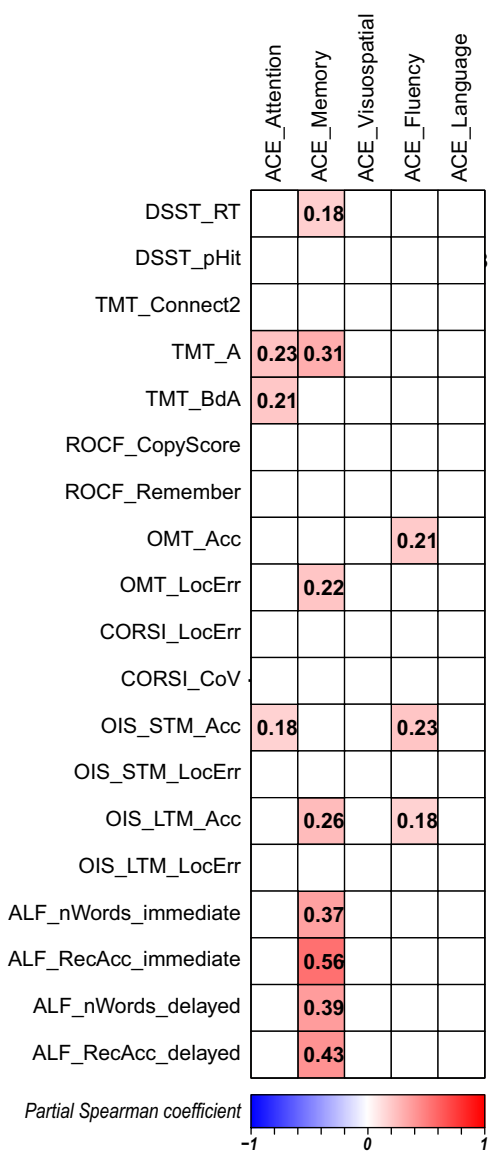
The final step was to establish whether OCTAL provides stable measurements over time, a prerequisite for monitoring cognitive trajectories or use in clinical trials. We therefore assessed 6-month test–retest reliability in both healthy participants and clinical groups, comparing OCTAL indices with repeated ACE-III administrations. A total of 118 participants drawn from the Study 3 cohort—59 cognitively healthy controls, 23 with SCD, 5 with MCI, and 31 with Alzheimer’s disease dementia—completed OCTAL on three occasions: at baseline (data presented in Study 3), 3 months, and 6 months (Table 4).

These participants additionally undertook the paper-based ACE-III using the alternate A, B and C forms. Test–retest reliability was quantified with the intraclass correlation coefficient (ICC) defined by McGraw and Wong<sup>21</sup>, which was calculated on the continuous performance scores across

**Table 3 | Correlations linking ACE-III and OCTAL metrics in Study 3 (N = 194)**

OCTAL Metrics	ACE-III Total		ACE-III Attention		ACE-III Memory		ACE-III Visuospatial		ACE-III Fluency		ACE-III Language	
	rho	adjusted p	rho	adjusted p	rho	adjusted p	rho	adjusted p	rho	adjusted p	rho	adjusted p
DSST_RT	0.55	<0.001	0.44	<0.001	0.50	<0.001	0.48	<0.001	0.44	<0.001	0.33	0.002
DSST_pHit	0.15	1.000	0.22	0.236	0.16	1.000	0.20	0.433	0.06	1.000	0.04	1.000
TMT_Connect2	0.49	<0.001	0.38	<0.001	0.42	<0.001	0.33	<0.001	0.41	<0.001	0.33	<0.001
TMT_A	0.59	<0.001	0.51	<0.001	0.59	<0.001	0.38	<0.001	0.50	<0.001	0.34	<0.001
TMT_BdA	0.50	<0.001	0.47	<0.001	0.47	<0.001	0.41	<0.001	0.41	<0.001	0.39	<0.001
ROCF_CopyScore	0.25	0.015	0.28	0.003	0.24	0.023	0.27	0.005	0.20	0.139	0.09	1.000
ROCF_Remember	0.41	<0.001	0.31	0.001	0.38	<0.001	0.26	0.012	0.36	<0.001	0.28	0.005
CORSI_LocErr	0.46	<0.001	0.38	<0.001	0.42	<0.001	0.40	<0.001	0.35	<0.001	0.31	0.001
CORSI_CoV	0.06	1.000	-0.02	1.000	0.06	1.000	-0.02	1.000	0.00	1.000	0.02	1.000
OIS_STM_Acc	0.48	<0.001	0.43	<0.001	0.40	<0.001	0.36	<0.001	0.46	<0.001	0.33	<0.001
OIS_STM_LocErr	0.13	1.000	0.05	1.000	0.14	1.000	0.00	1.000	0.09	1.000	0.10	1.000
OIS_LTM_Acc	0.58	<0.001	0.41	<0.001	0.55	<0.001	0.39	<0.001	0.50	<0.001	0.42	<0.001
OIS_LTM_LocErr	0.28	0.005	0.14	1.000	0.25	0.022	0.23	0.067	0.19	0.293	0.17	0.495
OMT_Acc	0.47	<0.001	0.42	<0.001	0.41	<0.001	0.37	<0.001	0.45	<0.001	0.35	0.002
OMT_LocErr	0.46	<0.001	0.31	0.011	0.46	<0.001	0.32	0.009	0.41	<0.001	0.33	0.007
ALF_nWords_immediate	0.74	<0.001	0.41	0.029	0.76	<0.001	0.53	<0.001	0.73	<0.001	0.50	0.002
ALF_RecAcc_immediate	0.61	<0.001	0.27	0.829	0.75	<0.001	0.44	0.011	0.59	<0.001	0.50	0.002
ALF_nWords_delayed	0.67	<0.001	0.35	0.147	0.69	<0.001	0.55	<0.001	0.60	<0.001	0.46	0.007
ALF_RecAcc_delayed	0.66	<0.001	0.41	0.033	0.71	<0.001	0.50	0.001	0.61	<0.001	0.46	0.006

All OCTAL variables were normalised against age-matched reference data; ACE-III scores are presented as raw scores, consistent with standard practice. Spearman correlation matrix relating the ACE-III total score and the subscores to OCTAL metrics. All p values have been adjusted for multiple comparison. Note that ALF refers to the Verbal Wordlist Recall task.



**Fig. 6 | Partial correlation matrix linking ACE-III subscores and OCTAL metrics in Study 3 (N = 194).** All OCTAL variables were normalised against age-matched reference data; ACE-III scores are presented as raw scores, consistent with standard practice. Partial correlations between each ACE-III subscore and the OCTAL metrics, calculated while regressing out the remaining four subscores (e.g. for ACE-Attention, memory, visuospatial, fluency, and language were removed as covariates). Note that ALF refers to the Verbal Wordlist Recall task.

the three visits. Following the interpretative framework of Koo and Li<sup>22</sup>, values below 0.50 were deemed poor, 0.50–0.74 as moderate, 0.75–0.90 as good and those exceeding 0.90 as excellent. All coefficients, ranked from lowest to highest, are displayed in Fig. 9, with their 95% confidence intervals and F-statistics reported in Supplementary Table 8.

The ACE-III total score demonstrated excellent stability over three tests (ICC = 0.98, 95% CI 0.97–0.98). Domain scores were likewise robust, with Memory 0.95, Fluency 0.92, Attention 0.91, Visuospatial 0.89 and Language 0.80; each coefficient significantly exceeded the 0.75 criterion for good reliability ( $p < 0.001$ ). It is worth noting, however, that the exceptionally high ICCs observed for the ACE-III total and for OCTAL’s DSST accuracy (DSST pHit, ICC = 1.0; see Fig. 9) are likely influenced by ceiling effects in unimpaired participants, where restricted variance inflates apparent test–retest stability.

Many OCTAL-derived metrics which do not have ceiling or flooring performance nonetheless surpassed these benchmarks despite a broader

performance range. The least stable measure, the Object-Memory Test location error (OMT LocError), still achieved an ICC above 0.75, denoting good reliability. these findings confirm that every principal OCTAL metric possesses good to excellent 6-month test–retest reliability, comparable with the ACE-III and its subscores.

## Discussion

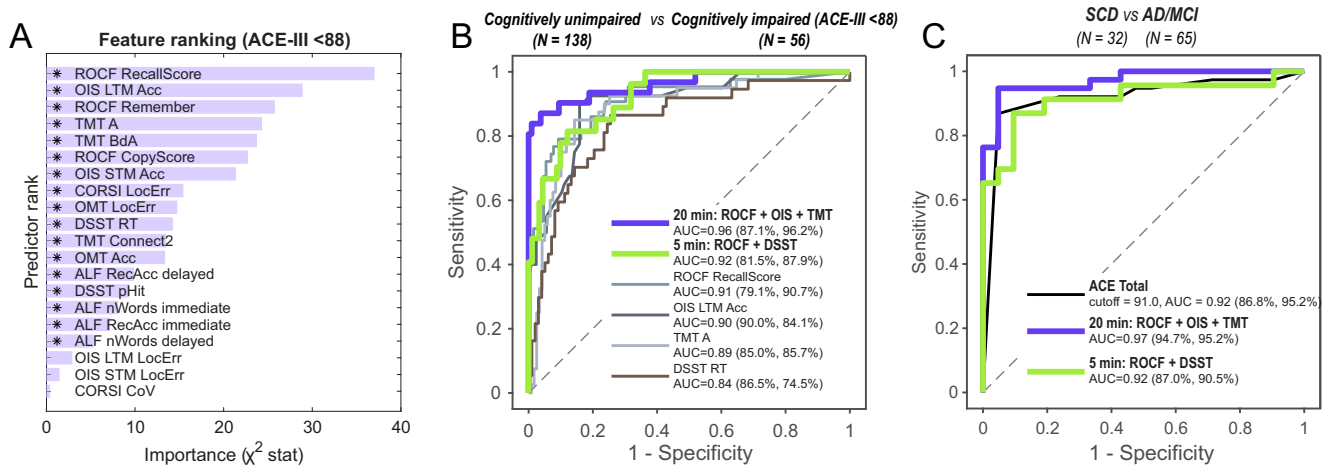
Here, we introduced the OCTAL, a remote, browser-based battery that measures memory, attention, visuospatial and executive functions (Fig. 1). Across four studies—spanning 1664 adults from young adults to patients with Alzheimer’s disease dementia, MCI or SCD—we demonstrated cross-cultural applicability, lifespan sensitivity, clinical utility, and robust reliability, demonstrating that OCTAL can serve as a scalable alternative to traditional pen-and-paper assessments. Crucially, patients with AD dementia and MCI completed OCTAL remotely without any clinician present, demonstrating its feasibility in cognitively impaired populations.

OCTAL builds upon a growing ecosystem of digital cognitive assessment tools, as reviewed recently in ref. 14, including established platforms like CANTAB<sup>23</sup>, the Cogstate Brief Battery<sup>24</sup>, and Cognitron<sup>19</sup>. A critical challenge for remote digital assessments, highlighted in large-scale studies with over 100,000 participants, is the difficulty in maintaining participant engagement<sup>25</sup>. With this challenge in mind, OCTAL was co-designed from the ground up by clinicians, cognitive neuroscientists, and patients (see ‘Methods’ for more details). This collaborative approach ensures that the platform not only meets clinical needs and scientific standards, but is also feasible for use in the patients with cognitive impairment. Future work will be required to assess and improve longitudinal retention in patient cohorts.

Another key advantage that distinguishes OCTAL is its open, modular architecture hosted on Pavlovia, a platform widely used by cognitive and psychological researchers<sup>26–28</sup>. This architecture facilitates straightforward integration with other web-based tools, such as Cognitron, and allows for the seamless incorporation of novel cognitive tests as they are developed. This inherent flexibility makes OCTAL a uniquely sustainable and evolvable tool for the research community.

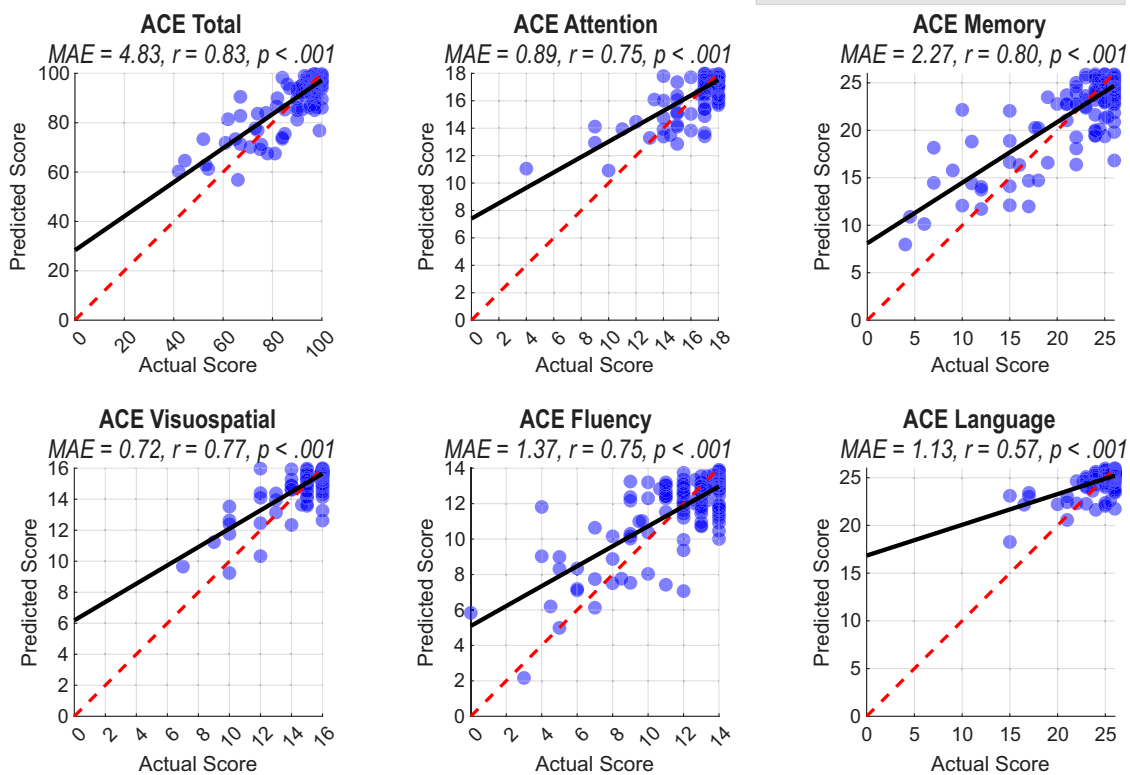
As shown in Study 3, the visuoconstructive recall (ROCF), delayed object–scene recognition (OIS) and visual processing speed (TMT-A) emerged as the strongest predictors of ACE-III-defined cognitive impairment (a standard face-to-face cognitive screening test) in the machine-learning analyses of 194 individuals (Fig. 7A). These top three tasks, taking 20 min in total, showed excellent performance for detecting cognitive impairment defined by an ACE-III cutoff of  $< 88$  (AUC = 0.96, Fig. 7B). The combination of ROCF and DSST also performed very well (AUC = 0.92) (Fig. 7B). Furthermore, the 20-min OCTAL subset outperformed the in-person ACE-III in distinguishing AD or MCI from subjective complaints (SCD), while an ultra-brief 5-min version (ROCF + DSST) matched ACE-III accuracy with markedly shorter testing time, given the fact that ACE-III typically takes 15–20 min (Fig. 7C). Furthermore, the modelling demonstrated that remote OCTAL performance could predict the specific scores for each ACE-III cognitive domain, showing a strong correlation with the actual clinician-administered scores for memory, attention, fluency, and visuospatial functions (Fig. 8). This highlights that OCTAL not only provides a valid global measure of cognition but also offers a granular, domain-specific profile that aligns closely with traditional, in-person assessments. While the present machine-learning results are encouraging, future work will extend these analyses to item-level feature behaviour to further test generalisability across populations.

OCTAL yields domain-specific measures of executive functions and memory that capture expected patterns of age-related variation and clinical impairment. This is demonstrated in Study 2’s investigation of healthy cognitive ageing, where OCTAL quantifies domain-specific change across five decades from mid to late life. Distinct cross-sectional age-related patterns were observed, with OCTAL revealing that executive functions decline from the mid-forties while memory remains within young-adult normative ranges until the mid-seventies. Long-term object–scene memory showed intermediate decline at  $\approx 55$  years (Fig. 2F), emphasising the value of



**Fig. 7 | OCTAL performance in detecting ACE-III-defined cognitive impairment and differentiating clinical cohorts.** A Univariate  $\chi^2$  feature ranking of OCTAL metrics for ACE-III defined cognitive impairment (i.e., ACE-III-Total score < 88) in study 3. Of 194 participants, 138 scored  $\geq 88$  (normal cognition) and 56 scored < 88 (cognitive impairment), irrespective of diagnosis or subjective symptoms. Higher importance scores denote greater predictive value; metrics are ordered accordingly, and \* marks statistical significance. Scores were converted to  $P$ -values using  $P = e^{-score}$ . B Receiver operating characteristic (ROC) curves illustrate the classification accuracy for cognitive impairment defined by ACE-III total score < 88. The area under the curve (AUC) values are shown for each model. The percentage values in brackets indicate sensitivity and specificity, respectively. C ROC analysis comparing OCTAL subsets with ACE-III Total for distinguishing participants with neurodegeneration (Alzheimer’s disease,  $n = 54$ ; mild cognitive impairment,  $n = 11$ ; total = 65) from those with subjective cognitive decline but no objective impairment ( $n = 32$ ). DeLong’s test shows the 20-min OCTAL subset marginally outperforms ACE-III Total ( $p = 0.04$ ), whereas the 5-min subset yields a comparable AUC.

**Predicted vs. Actual ACE-III subdomain scores**



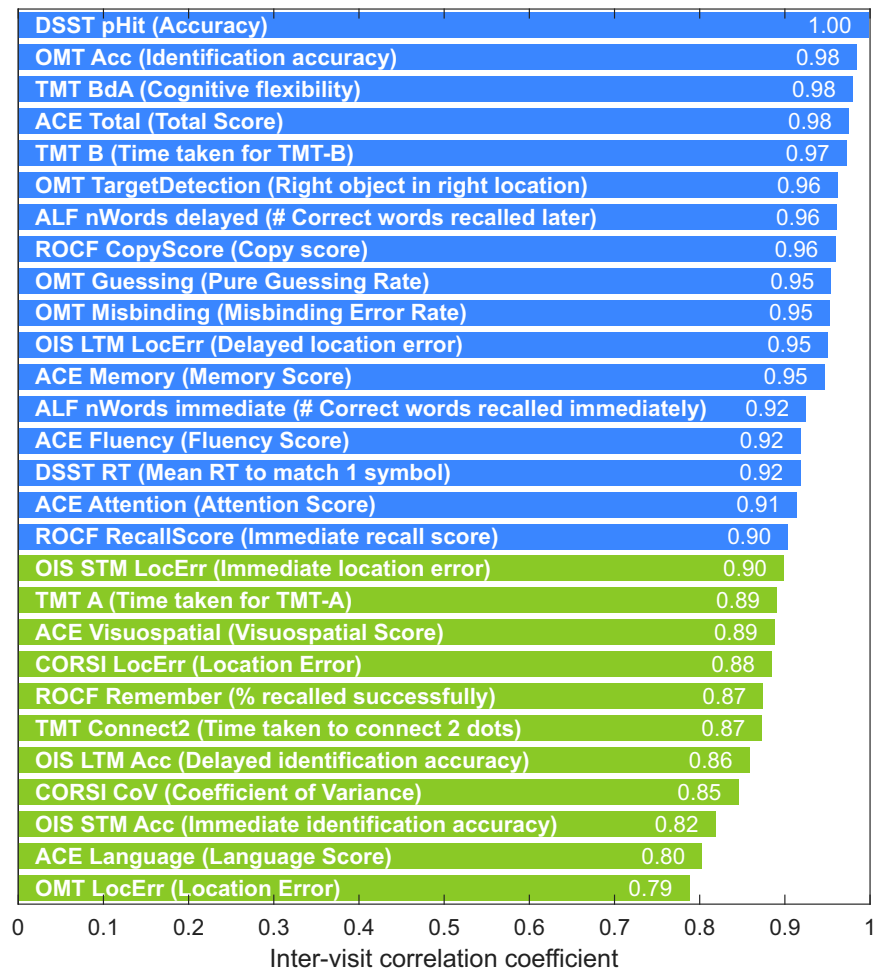
**Fig. 8 | Predicting ACE-III subscores from remote OCTAL performance in Study 3.** Each scatter plot compares scores on the ACE-III (total and subdomains) predicted by a multiple linear regression model against the actual, clinician-administered scores. The models used performance on the ROCF, OIS, and TMT tasks as predictors. Blue circles represent individual participants. The solid black line shows the linear best fit, while the dashed red line indicates perfect agreement ( $y = x$ ). Each panel is annotated with the Mean Absolute Error (MAE) and the Pearson correlation coefficient ( $r$ ) for that model.

**Table 4 | Demographics and ACE, OCTAL performance in four studies**

Study	Group	Total_N	Gender	Age	Education	ACE_Total	ROCF_RecallScore	OIS_LTM_Accuracy	OIS_LTM_LocErr	OMT_Acc	CORSI_LocErr	TMT_A	DSST_RT	ALF_nWords_delayed
1	English	194	F87	31.6 (6.9)	15.0 (2.0)	N.A.	92.4 (10.6)	77.9 (19.8)	0.2 (0.1)	91.4 (6.4)	5.4 (4.1)	20.3 (9.2)	2.4 (2.3)	8.4 (2.8)
			M105	[N = 192]	[N = 93]	[N = 184]	[N = 149]	[N = 120]	[N = 180]	[N = 149]	[N = 180]	[N = 149]	[N = 180]	[N = 149]
1	Chinese	167	F96	29.9 (5.7)	16.5 (1.5)	N.A.	95.1 (7.7)	81.3 (18.4)	0.2 (0.1)	94.4 (4.8)	N.A.	20.8 (7.1)	2.2 (0.5)	N.A.
			M71	[N = 167]	[N = 167]	[N = 167]	[N = 167]	[N = 167]	[N = 167]	[N = 167]	[N = 167]	[N = 167]	[N = 167]	[N = 167]
2	HC	1109	F626	54.1 (12.7)	15.3 (2.1)	N.A.	86.6 (13.7)	65.2 (23.0)	0.2 (0.1)	90.0 (7.4)	6.3 (4.5)	23.7 (9.0)	3.1 (2.6)	8.8 (2.8)
			M479	[N = 1106]	[N = 1109]	[N = 1037]	[N = 1037]	[N = 1037]	[N = 1037]	[N = 1037]	[N = 1037]	[N = 1037]	[N = 1037]	[N = 981]
3	HC	97	F53	60.7 (17.6)	17.9 (4.0)	97.1 (2.7)	86.5 (13.8)	64.7 (21.3)	0.2 (0.0)	91.3 (6.2)	6.5 (3.3)	27.4 (16.8)	3.0 (1.0)	9.0 (2.6)
			M44	[N = 97]	[N = 93]	[N = 97]	[N = 94]	[N = 90]	[N = 90]	[N = 97]	[N = 82]	[N = 82]	[N = 84]	[N = 61]
3	SCD	32	F12	61.4 (8.5)	15.9 (3.6)	95.4 (5.7)	88.5 (17.9)	74.4 (19.0)	0.2 (0.1)	89.5 (6.3)	7.5 (6.1)	26.9 (5.4)	3.7 (4.0)	8.5 (2.6)
			M17	[N = 29]	[N = 30]	[N = 32]	[N = 27]	[N = 27]	[N = 27]	[N = 23]	[N = 31]	[N = 27]	[N = 31]	[N = 31]
3	MCI	11	F4	67.0 (6.5)	12.6 (3.2)	88.5 (5.4)	72.3 (19.5)	57.9 (24.9)	0.2 (0.0)	84.1 (7.2)	10.3 (5.3)	35.5 (10.7)	4.2 (0.7)	6.7 (1.5)
			M6	[N = 10]	[N = 11]	[N = 11]	[N = 8]	[N = 8]	[N = 8]	[N = 8]	[N = 10]	[N = 10]	[N = 9]	[N = 8]
3	AD	54	F29	69.9 (8.9)	14.8 (4.6)	70.6 (16.7)	46.4 (28.0)	24.8 (19.3)	0.2 (0.1)	79.3 (10.1)	13.8 (7.5)	46.9 (20.3)	7.9 (6.2)	4.2 (2.9)
			M24	[N = 53]	[N = 47]	[N = 54]	[N = 43]	[N = 41]	[N = 41]	[N = 28]	[N = 47]	[N = 39]	[N = 35]	[N = 35]
3	Total	194	F98	63.7 (14.4)	16.5 (4.4)	88.9 (14.9)	76.1 (25.8)	56.1 (27.5)	0.2 (0.1)	87.5 (8.9)	8.9 (6.2)	32.5 (18.1)	4.5 (4.2)	7.1 (3.2)
			M91	[N = 189]	[N = 190]	[N = 194]	[N = 172]	[N = 166]	[N = 166]	[N = 116]	[N = 170]	[N = 159]	[N = 135]	[N = 135]
4	HC	59	F33	66.8 (11.6)	16.2 (3.2)	97.4 (1.9)	89.0 (11.6)	69.4 (17.8)	0.2 (0.0)	91.8 (4.7)	6.0 (2.9)	28.5 (10.2)	3.3 (1.1)	8.4 (2.9)
			M26	[N = 59]	[N = 56]	[N = 58]	[N = 59]	[N = 58]	[N = 58]	[N = 58]	[N = 57]	[N = 57]	[N = 58]	[N = 58]
4	SCD	23	F9	62.0 (9.0)	15.7 (3.4)	94.0 (4.2)	93.9 (7.8)	77.1 (17.9)	0.2 (0.0)	88.8 (5.7)	6.6 (3.1)	25.8 (5.2)	3.4 (1.8)	9.5 (1.9)
			M13	[N = 22]	[N = 22]	[N = 21]	[N = 23]	[N = 23]	[N = 21]	[N = 21]	[N = 23]	[N = 23]	[N = 22]	[N = 22]
4	MCI	5	F2	67.9 (4.7)	10.4 (0.5)	90.2 (4.5)	76.2 (17.8)	60.5 (17.3)	0.2 (0.0)	80.7 (6.0)	13.1 (7.0)	42.3 (8.5)	4.4 (0.8)	5.7 (1.5)
			M3	[N = 5]	[N = 5]	[N = 2]	[N = 5]	[N = 5]	[N = 5]	[N = 5]	[N = 5]	[N = 5]	[N = 5]	[N = 5]
4	AD	31	F18	70.1 (9.0)	14.3 (3.9)	72.2 (12.6)	43.1 (23.6)	21.6 (16.1)	0.2 (0.1)	79.2 (9.3)	14.3 (6.4)	50.3 (25.9)	11.0 (10.9)	4.5 (4.5)
			M12	[N = 30]	[N = 31]	[N = 24]	[N = 28]	[N = 30]	[N = 27]	[N = 27]	[N = 29]	[N = 27]	[N = 30]	[N = 30]
4	Total	118	F62	66.8 (10.5)	15.3 (3.6)	90.3 (13.0)	78.0 (25.3)	58.2 (27.9)	0.2 (0.1)	87.7 (8.2)	8.5 (5.6)	33.9 (17.8)	5.4 (6.5)	7.9 (3.4)
			M54	[N = 116]	[N = 110]	[N = 89]	[N = 113]	[N = 116]	[N = 116]	[N = 111]	[N = 114]	[N = 114]	[N = 111]	[N = 115]

The ACE-III and OCTAL metrics were averaged across assessments for Study 4.

**Fig. 9 | Test–retest reliability ranking.** Intraclass correlation coefficients (ICCs) are presented for all OCTAL metrics across three visits, alongside the ACE-III total and subdomain scores. ICCs were classified as poor (<0.50), moderate (0.50–0.74), good (0.75–0.90), or excellent (>0.90). Metrics with good reliability are coloured green; those with excellent reliability are coloured blue. Note that ALF refers to the Verbal Wordlist Recall task.



granular digital metrics over composite scores. These patterns are highly consistent with established findings from the cognitive ageing literature, where processing speed and executive functions show earlier and steeper declines than episodic memory<sup>29</sup>. Importantly, because our design was cross-sectional, these differences reflect population-level associations and should not be interpreted as within-person trajectories of decline. Longitudinal follow-up will therefore be required to confirm whether the observed associations translate into true age-related trajectories of cognitive change. Notably, the relatively weak inter-correlations between OCTAL metrics further suggest that the battery captures separable cognitive processes rather than collapsing onto a single latent *g* factor, again aligning with evidence that different domains of cognition exhibit distinct ageing profiles<sup>29</sup>. In psychometrics, the *g* factor refers to general intelligence, a latent construct thought to capture variance common to diverse cognitive tasks. Our exploratory analyses indicated that OCTAL tasks cluster into separable domains (executive vs memory), supporting the battery’s ability to deliver domain-specific profiles rather than a unitary global score. Importantly, performance on the visuomotor control task showed only modest latency increases across age, indicating that the observed executive decline cannot be explained solely by motor slowing or technological unfamiliarity.

Performance remained stable over 6 months (*Study 4*, ICCs ≥ 0.79), indicating good test–retest reliability. One possible explanation for the higher resolution observed with self-administered OCTAL compared with clinician-administered ACE-III is its avoidance of ceiling effects, which provides greater granularity in high-functioning individuals. That said, some accuracy-based tasks approached ceiling in unimpaired groups, a common feature of brief screening tools. Documenting maximum and minimum observed values in future datasets will clarify the extent to which

each task preserves sensitivity across the spectrum of ability. Other factors may also contribute. For instance, the digital, automated nature of OCTAL eliminates variability in test administration and scoring that can occur with a pen-and-paper test administered by different clinicians. Furthermore, the millisecond-level reaction time data captured by OCTAL offers a level of precision in measuring processing speed and executive function that is not possible with manually timed, observation-based assessments like the ACE-III.

A further significant advantage of this digital, automated nature is the ability to embed objective performance validity checks. While in the present study, we used these checks primarily to flag low-engagement data for exclusion, they offer a much richer opportunity. Performance on these embedded attention probes could be synthesised into a quantitative ‘engagement index’. For a clinician, such an index would provide a formal measure of test validity, replacing the subjective ‘soft features’—such as observing a patient’s effort or hesitation—relied upon during in-person assessments to gauge whether a cognitive score is a true reflection of the individual’s ability. This data-driven approach to measuring engagement is a key step towards more reliable and nuanced remote cognitive assessment.

Importantly, device heterogeneity did not diminish performance, as OCTAL ran reliably on any standard hardware without requiring a uniform screen size or input method (*Study 1*). Most OCTAL metrics—such as reaction time and spatial-error indices—were unaffected by screen size or aspect ratio (Supplementary Figs. 2 and 3), confirming its suitability for deployment on heterogeneous devices without formal standardisation. Where device effects appear (e.g. fine-motor precision in the ROCF copy task), OCTAL mitigates them by expressing memory performance as the recall-to-copy ratio, effectively regressing out motor variance. Expanding

normative datasets across device types may further refine this adjustment. Nevertheless, only one aspect of device heterogeneity was systematically tested here; future work will benchmark additional sources of variance (operating system, browser, hardware latency) that may interact with time-based metrics. Notably, PsychoJS/Pavlovia's timing precision has been systematically validated: large-scale benchmarking studies demonstrate millisecond-level accuracy in stimulus presentation and response logging across devices and browsers<sup>26,28</sup>, providing reassurance that device variability adds only modest noise relative to the cognitive effects under study.

Furthermore, cross-linguistic validation demonstrated equivalent performance between native English and native Chinese speakers once translations were implemented. The core metrics of the OCTAL were practically indistinguishable between the groups, falling within one standard deviation of the normative range (see Table 1). One likely contributor to the slower TMT-B performance observed in Chinese participants and other non-native English speakers is the added cognitive demand of alternating between numerals and English letters—a process that can be more effortful for second-language users. This effect may also reflect stimulus–response mapping complexity, whereby switching between symbol sets increases rule maintenance and set-shifting demands. While this explanation is plausible, a definitive test would require trial-level analyses of individual TMT steps to evaluate whether specific components of the task function equivalently across groups.

As OCTAL datasets expand, we will also develop language- and culture-specific normative references to complement pooled norms, improving interpretability in diverse populations. A somewhat surprising finding from our cross-cultural comparison was the consistent, albeit variable, latency in response times in the China-based cohort compared to the UK-based one. This was most pronounced for the initial identification response in the OIS task, but a general trend of slower reaction times was observed across tasks. This finding is important because, to our knowledge, detailed examinations of remote testing latency have been largely confined to North American and European internet environments<sup>26,28</sup>. The unique and complex nature of internet infrastructure and data handling in other regions, such as mainland China, has not been systematically documented in the context of remote cognitive testing. We do not consider this finding to be a (prohibitive) limitation for the use of time-sensitive measures in diverse online environments. Rather, it highlights the critical need for robust experimental designs that can account for and mitigate these baseline differences. One effective strategy is the use of within-task controls and ratio-based metrics. For example, instead of relying on the absolute completion time of TMT-B, the ratio of TMT-B to TMT-A can provide a purer measure of executive function by factoring out baseline psychomotor speed and general system latency. The observation that the most significant latency difference occurred in a rapid-response component of the OIS task, while being less pronounced in tasks with longer decision times, may suggest a larger initial latency at the beginning of a trial. This initial lag becomes a smaller proportion of the total reaction time in more complex tasks. This underscores the importance of moving beyond simply trusting that device and internet differences are negligible. Future research in remote cognitive assessment must prioritise study designs that are either validated across diverse technological landscapes or include internal controls to ensure that the derived cognitive metrics are robust and truly reflect cognitive processes, not environmental or technical confounders. This is an essential step for achieving genuine international scalability and equity in digital cognitive science. Screen-size analyses represent an important first step, but broader device heterogeneity and network environments will need to be incorporated routinely into validation pipelines.

Another example of cultural influence on cognitive performance is the Trail-Making Test Part B (TMT-B). Even when all participants lived in the UK (thereby reducing the likelihood that internet infrastructure differences explained the effect), healthy individuals who speak and read English fluently took considerably longer to complete this, while there was no difference in TMT-A. This pattern is consistent with the alphabet-switching demand of TMT-B imposing additional cognitive load on second-language

users, a finding consistent with previous reports in Korean speakers using the pen-and-paper version<sup>30</sup>. This finding underscores the long-standing challenge of ensuring the cross-cultural applicability of cognitive screening tools, a need that persists *even for fluent bilinguals*<sup>31–33</sup>. Although TMT-B is a highly sensitive and reliable predictor of cognitive impairment—surpassing all ACE-III subscores—its reliance on alphanumeric sequencing presents cultural barriers. Furthermore, its increasing duration with greater deficits poses practical challenges. Consequently, we omitted the TMT-B from the core smartphone version of OCTAL (still collecting data) to maintain both sensitivity and international usability. In its place, we incorporated the DSST, which also assesses attention and executive function, has a brief administration time of 2 min, and reliably predicts global cognitive impairment.

Together, these findings demonstrate OCTAL as a scalable, user-friendly and psychometrically robust alternative to traditional in-person cognitive assessment. Its domain-specific sensitivity, minimal participant burden and remote deployment render it ideally suited for large-scale cognitive screening, decentralised clinical trials and personalised longitudinal monitoring in precision neurology.

Moreover, remote cognitive testing has the capacity to fundamentally transform clinical-trial design<sup>11</sup>. By capturing high-frequency measurements, platforms like OCTAL can increase the statistical power of a study, which can in turn reduce the required sample size and associated costs. This approach also enables decentralised clinical trials, expanding participant access and making research in less common conditions more financially viable. Furthermore, by profiling individual cognitive trajectories rather than relying on group-mean effects, this level of precision supports adaptive trial protocols, accelerates 'go/no-go' decisions and facilitates truly personalised therapeutic interventions. Beyond clinical trials, these methods can benefit routine clinical services. The availability of rapid, low-burden assessments marks a significant shift away from infrequent, episodic testing that often requires specialist administration. This increased accessibility can streamline patient triage and help to inform more timely and effective care decisions<sup>14,15</sup>.

At the same time, several limitations should be acknowledged. First, the age-related findings reported here are based on cross-sectional comparisons. While such data are valuable for establishing expected patterns of cognitive ageing, they cannot be interpreted as within-person trajectories of decline. Future longitudinal follow-up of OCTAL cohorts will therefore be critical to determine whether the cross-sectional associations observed here translate into true age-related trajectories of cognitive change. Second, we did not assess the impact of colour-vision deficiencies on task performance. Although we deliberately avoided using colour in most tasks (ROCF, DSST, TMT and ALF) and designed OIS stimuli to differ in shape and orientation rather than hue—and varied fractal forms in OMT—participants with colour blindness may still experience slower visual discrimination. In Studies 3 and 4, colour blindness was one of the exclusion criteria. To address this in the future, we will include a brief screening question on colour vision in future studies. Third, we did not log the extent of caregiver assistance. Caregivers were instructed to clarify task instructions but not to perform the tasks for patients themselves. The strong concordance between OCTAL scores and ACE-III performance suggests that this instruction was generally followed, but formal monitoring would strengthen future protocols. Fourth, testing took place in uncontrolled home environments, without measures to limit external distractions or to standardise breaks between tasks. Although this reflects real-world conditions, it may introduce additional variance that should be quantified in subsequent work. Fifth, the current version of OCTAL does not include dedicated tests of language function, and none of the existing OCTAL tests correlated with the language subscore of the ACE-III. This might reduce the battery's sensitivity to isolated language deficits, such as those found in primary progressive aphasia. It should be noted, however, that the predominantly language-independent design is also a key strength, increasing the suitability of the battery for diverse populations where language proficiency could act as a confounding factor. Sixth, although all statistical comparisons were corrected for multiple testing, a

residual risk of Type I error cannot be entirely excluded; however, the convergence of results across independent cohorts and validation studies provides strong reassurance of the robustness of these findings. Finally, our data were pooled from multiple projects conducted at different time points, so not every participant completed every task; for example, the ALF task has a substantially smaller sample than the others.

Remote testing offers unprecedented opportunities for population health. It allows early detection of prodromal decline at scale, rapid assessments completed in under 10 min and, critically, unified integration with fluid biomarkers for refined risk stratification. The advent of scalable blood-based biomarkers for Alzheimer's disease necessitates digital tools that can be deployed remotely and longitudinally alongside them. In fact, in our previous study, we demonstrated strong relationships between OCTAL performance and plasma AD biomarkers (phosphorylated tau 181) in a smaller cohort<sup>34</sup>. Because no specialised hardware is required, OCTAL is accessible to geographically isolated or mobility-restricted individuals, and it supports long-term monitoring at intervals that would be impractical in a clinic setting. By enabling frequent, detailed sampling of cognitive function, OCTAL promises to accelerate precision medicine, enhance trial efficiency and reduce overall cost in both research and clinical practice.

## Methods

### Overview of the OCTAL platform

**Development and technical specifications.** The OCTAL (<https://octalportal.com>) is a remote, browser-based cognitive assessment platform designed for unsupervised administration on participants' personal devices, including desktops, laptops, and tablets. The platform's architecture supports a battery of cognitive tasks targeting diverse cognitive domains.

The majority of OCTAL tasks were developed using PsychoPy Builder (PsychoJS v2022.2.4), an open-source software package widely used for creating behavioural, psychological and cognitive experiments<sup>26,35</sup>. This choice of development environment promotes reproducibility and facilitates modification by the broader research community. To incorporate specific functionalities not natively available in PsychoJS, a custom-built in-house JavaScript library was developed and integrated. A notable exception to the PsychoJS framework is the ROCF task, which was custom-built from the ground up using HTML5 and JavaScript. A detailed discussion of the rationale and copyright considerations for the use of this stimulus is provided in the 'Ethical and intellectual property considerations' section. This approach was adopted to accommodate the complex drag-and-drop interactions required for its digital administration.

The development of OCTAL was led by a multidisciplinary team combining expertise in cognitive neuroscience, neurology, and software engineering (S.Z., M.H., S.To., S.G.M.). All task design and software implementation were completed in-house by S.Z., ensuring that experimental precision and clinical applicability were integrated from the outset. The platform was refined iteratively through internal discussions among clinicians and neuroscientists, together with informal usability feedback from patients and carers collected via S.To. during pilot testing. This feedback informed adjustments to task clarity, accessibility, and interface design. As this reflects standard patient and public involvement practice rather than a formal, separately documented study, details are summarised briefly here for conciseness.

This combination of standardised and bespoke development tools allowed for both efficient creation of common task elements and the flexibility needed for novel psychometric paradigms. The platform's open and modular architecture, a key design principle, facilitates straightforward integration with other web-based research tools and supports the future incorporation of new cognitive tests.

**Hosting, data anonymisation and security.** All OCTAL tasks are hosted on Pavlov.org, operating under the institutional licence of the University of Oxford's Department of Experimental Psychology. Pavlov.org is an established platform for hosting online psychological

experiments, providing robust infrastructure and data management capabilities. Throughout the data collection period detailed in this manuscript, OCTAL was deployed exclusively via this infrastructure. To support wider academic access and enhance scalability for future research, the platform is currently being transitioned to a dedicated server environment, also powered by Pavlov.org (<https://pavlov.org/octalportal.com/>).

Participant data are handled with strict adherence to privacy and security protocols. All collected data are fully anonymised; cognitive performance data are stored in an encrypted database, physically and logically separated from any personal identifying information. The Pavlov.org platform employs comprehensive security measures, including the use of EU-based servers that comply with stringent data protection regulations, such as the General Data Protection Regulation. All data transmission occurs over secure HTTPS connections utilising advanced encryption standards. Furthermore, access to the platform is blocked for browsers that do not support modern encryption technologies, ensuring a high level of data protection during remote assessment. These measures are critical for maintaining participant confidentiality and data integrity, particularly when collecting sensitive cognitive information from diverse populations, including clinical cohorts, in unsupervised remote settings.

**Usability and accessibility.** OCTAL was designed with a strong emphasis on usability and accessibility to ensure broad applicability. All tasks are delivered via standard web browsers, eliminating the need for participants to download or install any software, and no user log-in is required for participation. This approach minimises barriers to entry and simplifies the participant experience. The platform is compatible with a wide range of common web browsers, including Chrome, Firefox, Safari, and Edge, and is designed to function across various operating systems and devices, such as desktops, laptops, and tablets.

A core design principle was the optimisation of task presentation and interaction to ensure consistent performance and user experience regardless of variations in screen size or device type. This was partly addressed through specific analyses of screen size effects (detailed in Study 1) and the implementation of procedures like card calibration for tasks sensitive to stimulus dimensions (e.g. Oxford Memory Task). Critically, OCTAL was co-designed by a multidisciplinary team comprising clinicians, cognitive neuroscientists, programmers and patients with cognitive impairment. This collaborative, patient-centred design process was instrumental in creating a platform that not only meets rigorous scientific and clinical standards but is also engaging and intuitive for users, including older adults and individuals with cognitive difficulties. The task flow was deliberately optimised to be straightforward for these populations, addressing known challenges related to participant engagement and adherence in remote digital health studies.

**Task administration and flow.** To ensure a smooth and coherent testing experience, a custom JavaScript-based script manages seamless transitions between individual cognitive tasks within the OCTAL. This script automatically redirects participants from one completed task to the next, programmatically preserving a unique, anonymised participant identifier embedded within the URL. This system minimises the cognitive load on participants, minimises the data loss and reduces the likelihood of attrition between tasks.

The method for generating and managing participant links was adapted to different recruitment strategies. For large-scale online recruitment via platforms such as Prolific, the redirection and participant ID tracking were fully automated using Prolific's integrated participant ID system. For participants recruited from clinical settings (e.g. memory clinics), personalised study links were generated and sent via email following the informed consent process. These links embedded anonymised participant IDs and specific trial or visit identifiers, ensuring correct data attribution. Participants were generally encouraged to complete the entire battery of tasks in a single session to maintain consistency.

For researchers and clinicians, a secure backend interface, managed via GitLab, provided real-time access to monitor data completion and quality. This facility allowed for proactive identification of any technical issues or incomplete submissions. The system was designed to save incomplete data, with provisions for data deletion should a participant choose to withdraw consent, although no such requests were made during the studies reported.

**Ethical and intellectual property considerations.** The intellectual property associated with the OCTAL is formally managed by Oxford University Innovation (OUI), under project number 22768. Applications for a free academic licence can be made at: [octalportal.com/apply-academic-license](https://octalportal.com/apply-academic-license).

All visual stimuli and assets used within the OCTAL platform were either created for this project, are available in the public domain, or are used with permission from the copyright holders.

A specific consideration relates to our use of the ROCF. The stimulus for this task is the figure originally designed by Rey and later standardised by Osterrieth<sup>36,37</sup>. Our use of this figure is consistent with the principles of fair dealing for non-commercial academic research. Crucially, our novel digital adaptation employs only the figure as a visual stimulus; it does not use or replicate the copyrighted administration, scoring protocols, or manuals associated with the standardised pen-and-paper test. This digital task employs a drag-and-drop construction method, a procedure that is fundamentally different from the traditional pen-and-paper drawing administration. Despite this procedural difference, we have retained the 'ROCF' designation to reflect the task's primary aim: to assess visuospatial constructional ability, the same core cognitive domain targeted by the original test.

A core feature of this task's architecture is that it is intentionally stimulus-agnostic. This design allows for the future use of alternative, parallel stimuli, which is critical for mitigating practice effects in longitudinal studies where participants are assessed on multiple occasions. For the specific purpose of this initial validation study, however, we deliberately employed a single, fixed stimulus. The ROCF was selected for this benchmark role precisely because its well-characterised complexity and extensive documentation in the neuropsychological literature provide a stable and robust reference point. This approach ensures that our validation focuses on the performance of the digital platform itself, rather than introducing confounding variables from a novel or variable stimulus set.

### Task descriptions

The OCTAL comprises seven distinct tasks (Fig. 1), many adapted from established neuropsychological tests or behavioural paradigms, and optimised for robust remote administration across heterogeneous devices. These tasks are designed to measure various aspects of human cognition, including different forms of memory, attention, processing speed, and executive functions.

### Trail making test (TMT)

The TMT is a widely used neuropsychological test, adapted for online administration in OCTAL to assess attention, processing speed and executive functions, particularly cognitive flexibility<sup>38,39</sup>.

**Procedure.** Participants are instructed to tap or click on on-screen nodes (circled numbers or letters) as rapidly as possible to connect them in a specified sequence. The task includes three conditions:

- (1) Control Task ('Connect-2-Dots'): This condition consists of two trials. Two circles, labelled '1' and '2', are presented at opposite corners of the screen. Participants connect '1' to '2'. This task primarily gauges basic psychomotor speed and click accuracy.
- (2) TMT-A (Numeric Trails): This condition comprises three trials. Participants connect 25 circled numbers in ascending numerical order (e.g. 1-2-3-...).
- (3) TMT-B (Alternating Trails): This condition also comprises three trials. Participants connect 25 circled numbers and letters in an

alternating sequence (e.g. 1-A-2-B-3-C-...). For TMT-A and TMT-B, each participant is presented with six different trail maps (three for TMT-A, three for TMT-B), which are randomly selected from a pre-generated pool of 100 unique maps. These maps were created using a "divide-and-combine" algorithm to ensure consistent difficulty and spatial distribution<sup>40</sup>. The use of a large pool of distinct maps helps to minimise practice effects upon repeated administrations.

A demonstration of the TMT is available at: <https://octalportal.com/demo-trail-making/>.

**Outcome metrics.** The primary metrics derived from the TMT are mean completion times for each condition: TMT\_Connect2Dots (control), TMT\_A (numeric), and TMT\_B (alternating). Additionally, the ratio of TMT\_B completion time to TMT\_A completion time (TMT\_B/A ratio, also referred to as TMT\_BdA) is calculated as a specific marker of executive function, reflecting cognitive flexibility or task-switching ability. The digital format allows for millisecond-level precision in timing these responses

### Digit symbol substitution test (DSST)

The DSST is another commonly used measure of executive functions, processing speed, attention, and associative learning, adapted for digital administration within OCTAL<sup>41</sup>.

**Procedure.** A reference key is continuously displayed at the top of the screen, showing nine unique symbols, each paired with a digit from 1 to 9. Below this key, a row of nine randomised symbols is presented. Participants are required to match each symbol in the bottom row with its corresponding digit from the key by clicking on the correct digit (presented as a choice array). Once all nine symbols in the row are answered, the row is refreshed with a new set of nine randomised symbols. Participants are instructed to complete as many correct matches as possible within a fixed time limit of 2 min. A countdown in seconds is shown in the middle of the screen. Before the main task, two matched examples are shown, followed by a practice round with seven examples using a fixed practice key with 9 different symbols which do not appear in the main task. A demonstration of the DSST is available at: <https://octalportal.com/demo-dsst/>.

**Stimuli.** To mitigate practice effects, the main task's symbol-digit key is randomly generated for each session by selecting nine symbols from a pool of 20 distinct symbols. The rows of symbols for matching are generated dynamically. The symbols were purchased from the *Noun Project* (<https://thenounproject.com/>).

**Outcome metrics.** Performance is quantified by several metrics:

- **DSST\_rt:** The mean inter-response time for correct symbol-digit substitutions, calculated after removing outlier inter-response times (those more than 2 standard deviations from the mean of correct inter-response times). This is the key metric of DSST used in this paper.
- **DSST\_pHit:** The proportion of correct matches relative to the total number of responses made. This is the key metric of DSST used in this paper.
- **DSST\_nCorrectResponse:** The total number of correct symbol-digit matches completed within the 2-min duration.
- **DSST\_nAllResponse:** The total number of responses made by the participant during the task.
- **DSST\_totalIdleTime:** The cumulative duration of excessively long inter-response intervals for correct responses (defined as intervals greater than the median correct inter-response interval plus 3 times its standard deviation), reflecting periods of participant inactivity or disengagement during correct performance. This metric is used to exclude participants, not a cognitive measure.

### Freestyle Corsi Block Task (CORSI)

This task was inspired by the classic Corsi Block Tapping Task, a standard measure of visuospatial short-term memory span<sup>42</sup>. In the original version, participants were presented with a set of nine identical wooden blocks positioned on a board. Participants were required to point at the blocks in the order they were presented. On OCTAL, the key modification in our “Freestyle” version is that the spatial locations of the stimuli are not fixed across trials or participants, unlike typical computerised Corsi tasks where blocks appear in predefined positions. In an ‘*n*-location’ trial, a 1-cm wide red dot appears at a random location on the screen. After 1 s, it disappears. If the sequence length (*n*) is greater than 1, the dot then reappears at another random location for 1 s. This process repeats ‘*n*’ times, with sequence lengths tested up to 3 items. After the entire sequence of ‘*n*’ dots has been presented, there is a 1-s pause. Following this, the participant is instructed to reproduce the sequence by clicking on the screen at the locations where each dot had appeared, in the correct temporal order. The task is divided into three blocks. Each block consists of five trials of a specific sequence length (‘*n*-location sequence’). A demonstration of the CORSI is available at: <https://octalportal.com/demo-corsi/>

#### Outcome metrics:

- CORSI\_LocErr: The mean localisation error, calculated as the average Euclidean distance between the participant’s clicked response location and the actual target location for each item in the sequence, averaged across all trials and sequence lengths.
- CORSI\_CoV: The between-trial coefficient of variation of the localisation error, providing a measure of response consistency.

### Oxford memory task (OMT)

The OMT is a visual short-term memory task based on the “What was where?” paradigm, assessing memory for object identity and spatial location. This has been described in detailed in previous publications from our group (Fig. 2)<sup>43–45</sup>. This remote online version is shorten based previous data.

**Procedure.** Participants are presented with either one fractal pattern (easy trials) or three fractal patterns (hard trials) simultaneously displayed at various locations on the screen for 3 s (Encoding phase). Then, a 4-s delay was accompanied by a black screen (Maintenance phase). Following the delay, one of the previously presented fractal patterns (target) is shown alongside a novel fractal pattern (foil). The target and foil are displayed along the vertical meridian of the screen, with their relative positions (target above/below foil) randomised across trials (Identification phase). Participants must identify which pattern they just saw (identification performance) by clicking the target pattern and drag it to its proper location on the screen (localisation performance). A demonstration of the OMT is available at: <https://octalportal.com/demo-omt/>.

**Trials.** Participants begin with a practice block of 6 trials (3 one-item trials followed by 3 three-item trials). This is followed by a main test block of 40 trials, consisting of 20 one-item trials and 20 three-item trials. The order of trials within the main block is randomised for each participant. No performance feedback is given during either practice or main test blocks.

**Stimuli.** The stimuli are drawn from a library of 196 fractal images, comprising 49 distinct shapes, each available in 4 colour variations (<http://sprott.physics.wisc.edu/fractals.htm>). The foil pattern is drawn from the general pool of fractal images but is always one that was not presented as a target in the current trial (i.e. its specific shape and colour combination was not part of the to-be-remembered array for that trial). The on-screen locations for the fractal patterns during encoding are determined pseudo-randomly by an offline MATLAB script, with the constraint that fractals are never placed closer than 1.5 times of diameter to one another to avoid visual crowding and to ensure clear zones for localisation error analysis.

#### Outcome metrics:

- OMT\_Accuracy (Identification Accuracy): The proportion of trials where the target fractal is correctly identified. OMT\_Acc\_Easy means the mean identification accuracy for all easy trials (1 fractal). OMT\_Acc\_Hard means the mean identification accuracy for all hard trials (3 fractals).
- OMT\_LocError (Location Error): The Euclidean distance (e.g. in cm or pixels, then scaled) between the centre of the dragged fractal and the centre of its original target location. OMT\_LocErr\_Easy is the mean across all easy trials. OMT\_LocErr\_Hard is the mean across all hard trials.
- OMT\_IdERT (Identification Time): The reaction time from the presentation of the target-foil pair to the participant’s click identifying the target.
- OMT\_LocRT (Localisation Time): The reaction time from the initiation of the drag to the placement of the object at its final location.
- OMT\_TargetDetection: The rate of correctly identifying the target object AND placing it at the correct target location (a composite measure of accuracy).
- OMT\_Misbinding: The rate of correctly identifying a target but placing it at a location originally occupied by a different non-target item from the encoded array (in 3 fractal trials).
- OMT\_Guessing: The rate of placing an identified target randomly (“pure guess”), not corresponding to any of the original item locations.

### Rey–Osterrieth complex figure (ROCF)

The ROCF task in OCTAL is a digital adaptation of the traditional pen-and-paper test, designed to assess visuospatial perception, visuoconstructive abilities, and visuospatial episodic memory<sup>37</sup>. The original ROCF task requires the participant to draw a complex line-drawing freehand, first by replicating an existing figure (copy), and then again from memory (immediate recall). Our digitised version does not require hand drawing. Instead, the figure is split into 13 independent elements, and participants use a drag-and-drop interface to manipulate these elements on an empty digital canvas. The task comprises two phases:

- (1) Copy phase: Participants are shown the complete ROCF and are instructed to recreate it by dragging and dropping the predefined elements into their correct positions on the canvas. This phase assesses visuo-perceptual organisation and visuoconstructive skill.
- (2) Recall phase: Immediately following the copy phase, the model figure is removed, and participants are asked to reproduce the figure from memory, again by dragging and dropping the elements. This phase measures visuospatial episodic memory.

A demonstration is available at: <https://octalportal.com/demo-rocf/>

**Scoring.** An automated, offline scoring algorithm, implemented in MATLAB, is used to evaluate performance, providing a continuous measure of placement precision. This contrasts with the often more subjective, discrete scoring of the traditional version. The scoring logic is as follows:

- Anchor-based referencing: The large, central rectangle of the figure (‘a02\_large\_rectangle’) serves as the primary anchor element. The final (*x*, *y*) coordinates of all other placed elements are calculated relative to the centre of this anchor. These relative coordinates are then normalised by the width and height of the anchor element, respectively, to account for variations in the size of the constructed figure. If the primary anchor is not placed, the first element dropped on the canvas is used as a substitute anchor.
- Placement error precision: For each of the 13 elements placed on the canvas, the algorithm calculates the normalised absolute error in both the *x* and *y* dimensions. This error is the absolute difference between the element’s final normalised position and its ideal normalised position.
- Decay function: A ‘flat decay’ function is used to convert placement error into a score for each dimension (*x* and *y*). An element receives a

full score of 1 if its normalised error is within a tolerance of 0.1 (i.e. 10% of the anchor's dimension). For errors exceeding this tolerance, the score decays, approaching zero for large errors. This method provides credit for approximately correct placements while penalising significant deviations. Alternatively, the scoring function could also be linear function or exponential, where the score for an element decreases sharply and continuously with any deviation from its ideal location. Here, we chose to report the data with flat decay to be closer to the pen-and-paper ROCF scoring.

- Total score: The final score for each phase (Copy and Recall) is the sum of the *x*-scores and *y*-scores for all placed elements, scaled to a percentage (0–100%). The maximum possible score is reduced for each element that is not moved from the initial 'pile', effectively penalising omissions.

#### Outcome metrics:

- ROCF\_CopyScore, ROCF\_RecallScore: A proportional score (0–100%) reflecting the placement accuracy of the figure construction in the copy and recall phases, respectively.
- ROCF\_Remember: A memory retention index, calculated as the ratio of ROCF\_RecallScore to ROCF\_CopyScore (ROCF\_RecallScore / ROCF\_CopyScore \* 100%). This ratio normalises recall for individual differences in baseline visuomotor ability.
- ROCF\_CopyDuration: The total time in seconds from the start of the trial to its completion of the copy phase.
- ROCF\_RecallDuration: The total time in seconds from the start of the trial to its completion of the recall phase.

#### Object-in-scene memory task (OIS)

The OIS task is a newly developed memory paradigm, inspired by similar naturalistic object-in-photo tasks<sup>46</sup>, designed to simultaneously assess both short-term (tested immediately) and long-term (tested after a delay of over 1 min) visual associative memory. The task measures multiple components of memory for objects embedded within naturalistic scenes, including identification accuracy, precision of spatial localisation, and semantic memory aspects.

**Procedure.** Participants are presented with a photograph of a natural scene (e.g. a bamboo forest). A specific target object (e.g. a guitar) is shown placed at a random location within this scene. Participants are instructed to remember both the object and its location on the scene. To aid effective encoding, they are required to click on the displayed target object within the scene. Then, the memory is tested in two phases:

- Immediate recall (short-term memory): Immediately after encoding an object-scene pair, participants are presented with an array of 20 different objects. They must select the target object they just saw from this array and then drag and place it onto the original scene (now shown without the target object) at its remembered location. To make the object identification component more challenging and to assess semantic interference, the array of 20 objects includes a foil object that is semantically related to the target (e.g. if the target was a specific guitar, the foil might be a different type or colour of guitar).
- Delayed recall (long-term memory): After encoding and immediately testing a block of 5 different object-scene pairs, participants undertake a delayed recall phase. In this phase, they are probed on their memory for these 5 pairs, again requiring object identification and localisation. This delayed test typically occurs after an interval of ~1–2 min.

The task comprises a total of 20 trials (i.e. 20 unique object-scene pairs). These are typically divided into four blocks (e.g. 4 blocks of 5 pairs each for encoding and immediate recall, followed by delayed recall for those blocks). The order of object-scene pairs is randomised for each participant.

A demonstration is available at: <https://octalportal.com/demo-ois/>

**Stimuli.** The stimuli consist of photographs of everyday scenes and images of various common objects. The object stimuli used in this task were sourced from a standardised set developed by Brady et al.<sup>47</sup>.

**Outcome metrics.** Multiple metrics are extracted for both the immediate (STM) and delayed (LTM) recall phases:

- OIS\_STM\_Accuracy/OIS\_LTM\_Accuracy (Object Identification Accuracy): The proportion of trials in which the participant correctly identifies the original target object from the array. Chance performance for this metric is 5% (1 out of 20).
- OIS\_STM\_LocErr/OIS\_LTM\_LocErr (Location Error): The Euclidean distance, typically measured in centimetres on the calibrated screen, between the centre of the participant's placed object and the centre of the original target object's location in the scene.
- Semantic Identification Accuracy (OIS\_STM\_SemanticAcc/OIS\_LTM\_SemanticAcc): The proportion of trials in which the object identified by the participant belongs to the same semantic category as the target object, even if it is not the exact target item (e.g. choosing the foil guitar when the target was a different guitar). Chance performance for this is 10% (assuming two items from the same category are in the array, or as defined by the specific stimulus set construction).
- Identification Reaction Time (OIS\_STM\_IdRT/OIS\_LTM\_IdRT): Time taken to identify the object.
- Localisation Reaction Time (OIS\_STM\_LocRT/OIS\_LTM\_LocRT): Time taken to place the object. The OIS task's use of realistic scenes and objects aims to provide a more ecologically valid assessment of memory. The separation of object identity and spatial location memory, along with STM/LTM distinctions and semantic accuracy, allows for a comprehensive evaluation of visual associative memory.

#### Verbal wordlist recall (ALF)

This task is designed to assess verbal episodic memory, both reproduction and recognition memory, in immediate recall and delayed recall over 30 min delay. It was adapted from the Rey Auditory Verbal Learning Test (RAVLT) and from accelerated long-term forgetting (ALF) paradigms. In OCTAL documentation it has consistently been referred to as ALF; although the current implementation employs a short-delay adaptation (<1 h) rather than the conventional ≥24 h delay, we retained the abbreviation ALF to ensure continuity and consistency across OCTAL publications, software, and supporting materials.

**Procedure.** Participants are presented with a sequence of 12 words, presented one at a time. Each list contained one exemplar from 12 semantic categories (colour, animals, plants, vegetables, fruits, foods, materials, tools, clothing, parts of a building, weather, occupations), with assignment pseudo-randomised across a corpus of 100 unique word sets.

To promote intentional encoding and reduce variability in learning strategies, participants were explicitly instructed to study the words for a later memory test, mirroring procedures in standard verbal learning paradigms such as RAVLT.

To ensure adequate initial learning while still leaving room for subsequent forgetting, participants were required to reach a learning criterion of ≥50% correct recall (≥6 of 12 words) on the immediate recall test. This threshold was chosen to balance two goals: (i) minimising floor effects in delayed recall that can occur if participants fail to encode the list at all, and (ii) avoiding ceiling effects that limit sensitivity to decline in episodic memory performance. Similar encoding criteria are commonly used in list-learning tasks to ensure that performance at delay reflects retention rather than initial acquisition failure. Most unimpaired participants met this criterion on the first trial; the number of immediate recall attempts required was logged but not used in subsequent analyses.

- Immediate recall: Immediately after the presentation of the word list (and meeting the encoding criterion), participants engage in two tasks:
  1. Free recall: They are asked to type or verbally report (if proctored, though OCTAL is remote, so typing is assumed) as many of the 12 words as they can remember, in any order.

2. Recognition test: Following free recall, they complete a yes/no recognition test, where they are presented with words and must indicate whether each word was part of the list they just studied.
- Delayed recall: Approximately 30 min after the immediate test phase, during which other OCTAL tasks are administered, the memory test procedure is repeated. Participants again attempt free recall of the 12-word list and complete another yes/no recognition test for those words.

A demonstration is available at: [https://run.pavlovia.org/sijiazhao/demo\\_alf/](https://run.pavlovia.org/sijiazhao/demo_alf/)

**Stimuli.** The stimuli are lists of 12 common English words, selected to be of simple, high frequency. For scoring of free recall, only exact spellings of the target words are counted as correct (case insensitive). The corpus was selected carefully to make sure that the words are simple and basic (see Supplementary Table 9 for the full word corpus).

**Outcome metrics.** Performance on the ALF task is quantified by four primary metrics:

- ALF\_nWords\_immediate: The number of words correctly recalled (exact spelling) during the immediate free recall phase.
- ALF\_RecAccuracy\_immediate: Accuracy (proportion correct) on the immediate yes/no recognition test.
- ALF\_nWords\_delayed: The number of words correctly recalled (exact spelling) during the delayed free recall phase (after the 30-min interval).
- ALF\_RecAcc\_delayed: Accuracy on the delayed yes/no recognition test.

This task allows for the assessment of initial learning, ensured that the participants, susceptibility to interference over a delay, and the distinction between recall and recognition memory processes. The use of semantically categorised lists also offers the potential for analysing recall strategies (e.g. semantic clustering), though this is not explicitly mentioned as an outcome metric in the provided text.

### Study designs and procedures

The overview of all four studies' demographics and in-person ACE-III and key OCTAL metrics mean can be found in Table 2.

**Attention check.** To ensure data quality from unsupervised remote testing, several task-specific attention checks were embedded for all healthy controls: (1) OIS: Participants were required to drag an object onto the scene immediately after viewing; object identification accuracy <20% (where chance for correct object was 10% and for correct semantic category was 20%) was flagged. No participants failed this check. (2) OMT: An unreasonably short localisation time (cut-off at 0.2 s) was flagged. Two participants failed this check. (3) DSST: A correct rate <20% (chance for pure guessing ~11%) or being idle for more than 60 s in total was flagged. No participants failed these checks. (3) ROCF Copy: A copy score <80% was flagged. One participant failed this check. (4) Participants who failed only one attention check were retained in the dataset; however, their performance data for the specific task on which the attention check was failed were discarded from analyses involving that task. This strategy balanced the need for data quality with the desire to retain as much valid data as possible. Feasibility was therefore assessed through embedded attention and validity checks, which confirmed within-task engagement.

### Study 1

This study aimed to assess the feasibility and international scalability of the OCTAL platform and to determine whether its cognitive tasks exhibited any inherent bias towards English speakers when administered to young adults from different linguistic backgrounds.

**Participants.** A total of 361 young adults, aged 18–44 years, participated. The sample was divided into two cohorts: (1) English-speaking

Cohort:  $N = 194$  native English speakers residing in the United Kingdom. Recruitment was conducted via the Prolific online platform. (2) Chinese-speaking Cohort:  $N = 167$  native Chinese speakers residing in mainland China. For a sub-analysis on the effects of screen size, an additional 43 adults aged 45–70 from mainland China were included, expanding this specific analysis cohort to  $N = 210$ . Recruitment was conducted via the Credamo online platform.

Demographic information, including age, gender, and years of education, was collected. For the Chinese cohort, reported education levels were converted to an equivalent years-of-education scale to facilitate cross-cultural comparisons (e.g. 'Primary school and below' =  $\leq 6$  years, 'Bachelor's degree' = 16 years, 'Doctoral degree' = 23 years). The education level options, and year conversions are listed below:

- '小学及以下' (Primary school and below): 6 years or less
- '初中' (Junior high/Middle school): 9 years (6 + 3)
- '普高/中专/技校/职高' (High school/Vocational schools): 12 years (6 + 3 + 3)
- '专科' (Junior college): 15 years (6 + 3 + 3 + 3)
- '本科' (Bachelor's degree): 16 years (6 + 3 + 3 + 4)
- '硕士' (Master's degree): 18–19 years (16 + 2-3): 19 (using the higher estimate)
- '博士' (Doctoral degree): 21–23 years (18-19 + 3-4): 23

**Procedure.** All participants completed the OCTAL battery remotely on their own laptop or desktop computers. The English version of OCTAL was administered to the UK cohort, while a fully translated Chinese version was administered to the China cohort. For the screen size analysis, tablets were disallowed, and the browser automatically captured screen width and height in pixels at the time of testing.

### Study 2

This study aimed to evaluate the concurrent validity of remote OCTAL assessments by comparing performance with the ACE-III, an established in-person cognitive screening tool.

**Participants.** A large cohort of 1109 UK residents was recruited via Prolific.co. Of these, 905 participants were aged between 45 and 89 years, forming the primary sample for the cognitive aging analyses. Participants self-reported as neurologically healthy, with normal or corrected-to-normal vision, and no colour blindness. To minimise expectancy effects, the study was advertised as a 'brain game,' blinding participants to its specific cognitive assessment focus. This sample incorporated 352 healthy adults aged over 50 from a previously published normative study. Recruitment occurred in successive batches to achieve a broad age distribution (overall range 18–85 years), with particular emphasis on obtaining a balanced sample above age 40.

**Procedure.** All assessments were completed remotely and without supervision, using participants' personal devices. Participants were informed prior to starting that attention checks were embedded within the tasks and that low-effort responses might lead to data rejection.

### Study 3

This study aimed to evaluate the concurrent validity of remote OCTAL assessments by comparing performance with the ACE-III, an established in-person cognitive screening tool.

**Participants.** A total of 194 individuals participated, comprising diverse clinical and control groups: 97 cognitively healthy controls (HC), 54 patients with early Alzheimer's disease dementia (AD), 11 patients with MCI, and 32 individuals with SCD. Patients with AD were recruited from the Cognitive Disorders Clinic at the John Radcliffe Hospital, Oxford, UK. They presented with a progressive, multidomain, predominantly amnesic cognitive impairment, and had undergone Magnetic Resonance

Imaging (MRI) and Fluorodeoxyglucose Positron Emission Tomography (FDG-PET) imaging consistent with a clinical diagnosis of AD dementia (e.g. temporo-parietal atrophy and hypometabolism)<sup>48</sup>. All AD, MCI and SCD patients had plasma biomarkers examined, including pTau217, which were in line with their clinical diagnosis. MCI and SCD patients were diagnosed according to Petersen and Jessen's criteria<sup>49,50</sup>. HC were recruited from the same clinic as spouses of patients or through open day events. All HCs were aged >50 years, reported no psychiatric or neurological illnesses, were not taking regular psychoactive medications, scored above the cut-off for normality on the ACE-III ( $\geq 88/100$ ), and had a normal brain MRI scan as reviewed by two independent senior neurologists (S.T. and M.H.). Participants with colour blindness were excluded at the recruitment stage.

**Procedure.** (1) OCTAL assessment: Participants completed the OCTAL battery remotely on a personal desktop, laptop, or tablet computer using Chrome or Firefox browsers. A unique, anonymised link containing participant and visit identifiers was emailed to them on the same day as their in-person clinic visit. Participants were encouraged to complete the online tests within 1 week of the visit. (2) ACE-III assessment: All participants completed the ACE-III in person during their clinic visit, administered by trained personnel as a standard clinical cognitive screening measure. An ACE-III total score  $< 88/100$  was considered indicative of cognitive impairment. While HCs scoring below this cut-off were excluded, patients with AD were included based on clinical and radiological criteria, not solely on their ACE-III score. MCI was defined as ACE-III total score between 87/100 and 83/100.

The order of administration (remote OCTAL vs. in-person ACE-III) was pseudo-balanced across the cohort to control for potential order effects. This design allowed for a direct comparison of a novel remote digital assessment with a traditional, clinician-administered gold-standard screening tool across a spectrum of cognitive abilities.

#### Study 4

This study aimed to quantify the test-retest reliability of the OCTAL battery over a 6-month period.

**Participants.** This sample was drawn from the same recruitment pool as Study 3. A cohort of 118 participants was included, representing a similar spectrum of cognitive statuses as in Study 3: 59 cognitively healthy controls (HC), 23 individuals with SCD, 5 with MCI, and 31 with Alzheimer's disease dementia (AD).

**Procedure.** (1) OCTAL Assessments: Participants completed the full OCTAL battery on three separate occasions: at baseline, at 3 months post baseline, and at 6 months post baseline. (2) ACE-III Assessments: A subset of 89 participants also completed the paper-based ACE-III at each of these three time points. To minimise practice effects on this comparator measure, alternate forms of the ACE-III (Forms A, B, and C) were used across the assessments. This longitudinal design is essential for establishing the stability of OCTAL metrics over time, a critical property for tools intended for longitudinal monitoring or use in clinical trials. The inclusion of various clinical groups tests the reliability across different levels of cognitive function.

**Ethics.** Ethical approval for studies involving UK-based participants was granted by the University of Oxford ethics committee (IRAS ID: 248379, Ethics Approval Reference: 18/SC/0448). For Study 1, ethical approval for the recruitment of Chinese participants was granted by the Medical Sciences Interdivisional Research Ethics Committee (MS IDREC) at the University of Oxford (Reference: 671456). All studies were conducted in accordance with the Declaration of Helsinki, and all participants provided written or digital informed consent prior to starting any procedures.

## Statistical analysis

### General approach

Statistical analyses were performed using MATLAB (version R2025a; MathWorks, Inc.), R (version 12.0; R Core Team), and JASP (version 0.16.4; JASP Team)<sup>51</sup>. Unless otherwise specified, a two-tailed  $p$ -value  $< 0.05$  was considered statistically significant. Bonferroni correction for multiple comparisons was applied where appropriate (e.g. for large correlation matrices or multiple group comparisons on various metrics) to control the family-wise error rate. Effect sizes, such as the rank-biserial correlation ( $r$ ) for Mann-Whitney U tests, were reported to provide information on the magnitude of observed differences between groups.

### Specific analyses for Study 1

**Group comparisons.** Demographic characteristics and OCTAL performance metrics were compared between the English-speaking and Chinese-speaking cohorts using Mann-Whitney U tests, suitable for two independent groups, particularly when normality of distributions cannot be assumed. Test statistics ( $U$ ), effect sizes ( $r$ ), and  $p$ -values were reported. Beyond statistical significance, the practical significance of differences in OCTAL metrics between the two language groups was evaluated using a bootstrap permutation procedure with 10,000 iterations. For each metric, the mean performance of the Chinese group was compared against a normative range defined as  $\pm 1$  standard deviation (SD) from the mean of the English-speaking cohort. A difference was considered practically significant if  $\geq 97.5\%$  of the resampled means from the Chinese group fell below the lower threshold (English Mean - 1 SD) or above the upper threshold (English Mean + 1 SD), corresponding to a one-sided  $p < 0.025$ . This provides a more robust interpretation of group differences relative to normative variation.

**Screensize correlations.** To investigate the potential influence of display dimensions on OCTAL performance in the remote testing environment, Spearman rank correlations ( $\rho$ ) were calculated between screen width (and separately, screen height), captured in pixels, and all OCTAL metrics for the expanded Chinese cohort ( $N = 210$ ).  $P$ -values for these correlations were Bonferroni-corrected for the number of metrics tested. To formally compare the strength of correlations (e.g. to test if screen width and screen height had significantly different associations with ROCF Copy score), Hittner, May, and Silver's (2003) modification of Dunn and Clark's (1969)  $z$ -test for dependent overlapping correlations was used, as implemented in the `COCOR` package in R<sup>52-54</sup>.

### Specific analyses for Study 2

**Group comparisons.** Differences in demographic variables between subgroups (e.g. native vs. non-native English speakers) were assessed using Mann-Whitney U tests for continuous variables and chi-squared ( $\chi^2$ ) tests for categorical variables.

**Exploratory factor analysis (EFA).** To investigate the underlying structure of cognitive performance as measured by OCTAL, an EFA was conducted on 19 key OCTAL metrics from the large healthy cohort ( $N = 1109$ ). The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's Test of Sphericity ( $\chi^2$ ) were used to confirm the suitability of the data for factor analysis. A KMO value of 0.68 indicated acceptable sampling adequacy, and a significant Bartlett's test supported the factorability of the correlation matrix. Factors were extracted based on eigenvalues  $> 1$ . An initial analysis used varimax rotation (orthogonal). To allow for potential correlations between cognitive factors, the analysis was repeated using promax rotation (oblique). A  $\chi^2$  test assessed the sufficiency of the identified number of factors. This EFA helps to understand the main cognitive constructs captured by the OCTAL battery.

**Age trajectories.** To visualise age effects across the lifespan, we applied a year-by-year local smoothing approach. (1) Normalisation: Individual

participant scores on each OCTAL metric were standardised by converting them to  $z$ -scores relative to the mean and standard deviation of a young adult reference group (aged 18–39 years, derived from Study 1 or the younger participants in Study 2). Prior to normalisation, outliers defined as scores  $>2$  SD from the weighted mean at each age point were removed from this specific trajectory analysis to ensure robust trend estimation. The formula used for normalisation was (see Eq. (1)):

$$z = \frac{X - \mu_{\text{young}}}{\sigma_{\text{young}}} \quad (1)$$

where  $X$  represents the OCTAL outcome metric raw value,  $\mu_{\text{young}}$  and  $\sigma_{\text{young}}$  represent the mean and standard deviation of the young norm, respectively. For reaction time variables and location error variables,  $z$ -scores were sign-reversed so that more negative values consistently indicated poorer performance (i.e. greater decline). (2) Smoothing: Year-by-year age trajectories for these normalised scores were generated using a local smoothing technique. Performance at each target age was estimated as a Gaussian-weighted average of participants within a  $\pm 5$ -year window of that age. The weighting function was a Gaussian kernel, defined by a bandwidth ( $\sigma$ ) of 2.5 years (see Eq. (2)):

$$W(x) = e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2)$$

Here,  $x$  is the participant's age and  $\mu$  is the target age at the centre of the smoothing window. (3) Linear fits: To compare the rates of age-related change across different cognitive domains and metrics, straight lines were fitted to the Gaussian-smoothed, age-weighted mean  $z$ -scores for each metric (from ages 45 to 75 years) using the `polyfit` function in MATLAB. The slopes of these lines provided an estimate of the average annual change in performance (in  $z$ -units per year) and  $R$  squared measured the fitness of the linear model (reported in Supplementary Table 5).

**Inter-metric correlations.** Spearman rank correlation matrices were computed across all key OCTAL metrics to examine the relationships between different tasks and cognitive domains within the healthy adult sample. Results were presented both with and without Bonferroni correction for multiple comparisons to show the full pattern of relationships and those robust to stringent correction. Correlations between age and individual OCTAL metrics were also calculated using Spearman's  $\rho$ .

**Network module detection.** To investigate the underlying structure of cognitive performance, we constructed a weighted, undirected network from 25 cognitive metrics. The nodes in this network represented each metric, and the edge weights between them were defined by the absolute value of the pairwise Spearman correlation coefficient. We then applied a single-pass Louvain-style modularity maximisation algorithm to partition the network into distinct functional communities, or modules.

The Louvain algorithm iteratively optimises network modularity,  $Q$ , to find the best partition. Modularity is a scalar value between  $-0.5$  and  $1$  that measures the density of connections within modules compared to connections between modules.  $Q$  is formally defined as Eq. (3):

$$Q = \frac{1}{2m} \sum_{ij} \left[ A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \quad (3)$$

where  $A_{ij}$  represents the weight of the edge between nodes  $i$  and  $j$ ,  $k_i$  and  $k_j$  are the sum of weights (strength) of all edges attached to nodes  $i$  and  $j$  respectively,  $m$  is the sum of all edge weights in the network ( $m = \frac{1}{2} \sum_{ij} A_{ij}$ ), and  $\delta(c_i, c_j)$  is 1 if nodes  $i$  and  $j$  belong to the same community ( $c_i = c_j$ ) and 0 otherwise. The algorithm iteratively moves nodes between communities to maximise the gain in  $Q$ , and a full pass through all nodes is repeated until no further increase in  $Q$  can be achieved.

To confirm the stability of the detected community structure, we ran the Louvain algorithm for 10,000 iterations. In each iteration, the

algorithm's internal random node ordering allowed for exploration of different local maxima. The modularity  $Q$  for each resultant partition was computed using the formula above. The 6-module solution was robustly obtained in 83.99% of these runs ( $Q = 0.13 \pm 0.00$ ), confirming the high stability of this partition.

Finally, to identify the most influential metric within each community, we calculated the intra-modular strength for every node—defined as the sum of its connection weights to all other nodes within the same module—and designated the node with the highest strength as the central node for that community.

**Language effect analysis.** To explore whether having English as a second language affected OCTAL performance across the lifespan, a non-parametric bootstrap-based resampling method (2000 iterations) for independent samples was employed. This analysis compared performance between native English speakers and non-native English speakers at each age point, identifying age intervals where significant group differences emerged (defined as  $p < 0.05$ , two-tailed, uncorrected for this exploratory analysis).

### Specific analyses for Study 3

**Age-adjustment of OCTAL metrics.** For comparisons with clinical diagnosis classifications and cognitive impairment defined by ACE-III scores, each participant's OCTAL metrics were age-adjusted. Each individual's score was adjusted based on the mean and SD of normative participants from Study 1 and Study 2 within a  $\pm 3$ -year age band around their own age. On average, each individual's performance was adjusted against 55.8 normative participants.

**Correlations between OCTAL and ACE-III.** Spearman rank correlations were used to assess the association between age-adjusted OCTAL metrics and the ACE-III total score. A Bonferroni-corrected significance level of  $\alpha = 0.0026$  was applied for these correlations. To investigate domain-specific relationships, partial Spearman correlations were computed between each OCTAL metric and each of the five ACE-III subscores (Attention, Memory, Fluency, Language, Visuospatial), while statistically controlling for the variance accounted for by the other four ACE-III subscores. This helps to identify unique contributions of OCTAL tasks to specific cognitive domains as defined by the ACE-III.

**Machine learning for feature ranking.** To identify which OCTAL metrics were most predictive of cognitive impairment (defined as an ACE-III total score  $< 88$ ), a univariate feature ranking approach was used. For the binary classification (impaired vs. unimpaired), chi-square tests (`fschic2` function in MATLAB) were employed to assess the dependence of each OCTAL metric (feature) on the group classification. The resulting importance scores ( $-\log(p)$ ) were converted to  $p$ -values using  $P = e^{-score}$ . For predicting a continuous response variable like the ACE-III total score,  $F$ -tests via `fsrftest` in MATLAB was used.

**Receiver operating characteristic (ROC) analysis.** ROC analyses were conducted to evaluate the ability of individual OCTAL metrics and combinations of metrics to discriminate between cognitively impaired (ACE-III  $< 88$ ) and unimpaired (ACE-III  $\geq 88$ ) individuals. Area Under the Curve (AUC), sensitivity, specificity, and optimal cut-off thresholds (for both raw scores and age-adjusted  $z$ -scores) were determined for key OCTAL metrics and for pre-defined task combinations (e.g. a 20-min OCTAL subset: ROCF, OIS, TMT; and a 5-min OCTAL subset: ROCF, DSST).

**Comparison of ROC curves.** DeLong's test was used to statistically compare the AUCs of different diagnostic models (e.g. to compare the AUC of the 20-min OCTAL subset against the AUC of the ACE-III total score for distinguishing AD/MCI patients from individuals with SCD)<sup>55</sup>.

**Predicting ACE-III scores from OCTAL metrics.** To determine the extent to which remote OCTAL performance could predict scores on a standard clinical assessment, we conducted a series of multiple linear regression analyses using MATLAB (R2025a, MathWorks, Inc.) with the Statistics and Machine Learning Toolbox.

The predictor variables consisted of a subset of nine age-adjusted OCTAL metrics: time to connect two dots (TMT\_Connect2), time taken for TMT-A (TMT\_A), the TMT B/A ratio (TMT\_BdA), ROCF copy and remember scores (ROCF\_CopyScore, ROCF\_Remember), and object identification accuracy and localisation error for both short- and long-term memory in the OIS task (OIS\_STM\_Acc, OIS\_STM\_LocErr, OIS\_LTM\_Acc, OIS\_LTM\_LocErr). The outcome variables were the ACE-III total score and its five subscores (Attention, Memory, Visuospatial, Fluency, and Language).

A separate multiple linear regression model (fitlm function in MATLAB) was constructed for each of the six ACE-III outcomes. For each model, participants with incomplete data for any of the included predictor variables or the specific outcome variable were excluded from the analysis on a pairwise basis. The predictive accuracy of each model was assessed by correlating the model-predicted scores with the participants' actual scores; Pearson's  $r$  was used because it directly quantifies the strength of the linear relationship between the predicted and observed scores, which aligns with the primary objective of a linear regression model to establish the best linear fit to the data. The Mean Absolute Error (MAE) was calculated to quantify the average prediction error.

**Multi-cutoff classification and multivariate feature selection.** To test the robustness of feature importance and classification performance, we extended the analyses beyond the initial  $\chi^2$ -based univariate ranking. First, we repeated all classification analyses at three established ACE-III thresholds for impairment: <88, <85, and <82<sup>20</sup>. This allowed evaluation of stability across the cut-offs commonly used in clinical and research settings.

Second, to account for potential multivariate structure and redundancy among OCTAL metrics, we implemented two complementary machine learning approaches: penalised logistic regression (LASSO) and Random Forests (RF). LASSO regression was performed with nested tenfold cross-validation, including standardisation within the training folds to avoid information leakage. Model performance was summarised as mean area under the receiver-operating-characteristic curve (AUC) across folds, with 95% confidence intervals estimated by bootstrapping. Feature importance was indexed by the absolute magnitude of non-zero coefficients and selection frequency across resamples. Random Forests (400 trees, balanced class weights) were trained using out-of-bag (OOB) samples for unbiased performance estimation. Feature importance was derived from permutation importance scores, reflecting the reduction in predictive accuracy when each feature was randomly shuffled.

Stability analyses were conducted at two levels. (i) Within-cutoff resampling stability was quantified as the mean pairwise Spearman correlation ( $\rho$ ) between feature-importance vectors across repeated resamples of the same cutoff. (ii) Cross-cutoff stability was evaluated as the Spearman correlation between feature-importance rankings obtained at <88, <85 and <82 within each model set (20-min vs. 5-min composites).

#### Specific analyses for Study 4

**Age-adjustment of OCTAL metrics.** For longitudinal analyses in Study 4, OCTAL scores were age-adjusted using the same procedure as in Study 3. At each timepoint, an individual's score was adjusted based on the mean and SD of normative participants from Studies 1 and 2 within a  $\pm 3$ -year age band around their own age. On average, each participant's performance was referenced against 55–60 normative individuals per band. This approach ensured comparability of scores across baseline, 3-month, and 6-month assessments, while controlling for expected age-related differences in performance.

**Test-retest reliability.** The primary analysis for assessing the 6-month test-retest reliability of OCTAL metrics and ACE-III scores was the ICC, originally defined by McGraw & Wong<sup>21</sup>. ICC values were interpreted according to a more recent framework proposed by Koo & Li: values < 0.50 were considered poor, 0.50–0.74 as moderate, 0.75–0.90 as good, and >0.90 as excellent reliability<sup>22</sup>. ICCs, along with their 95% confidence intervals and associated F-statistics, were computed using MATLAB function ICC<sup>56</sup>. The F-statistics specifically tested the null hypothesis that the true ICC was equal to 0.75 (the threshold for "good" reliability), allowing for a rigorous assessment against this benchmark. This comprehensive approach to reliability assessment is vital for tools intended for repeated measures or longitudinal tracking.

#### Data availability

Normative data for OCTAL, stratified by age, are publicly available on GitHub (<https://github.com/sijiazhao/octal-normative-data>)(<https://github.com/sijiazhao/octal-normative-data>). This repository will be regularly updated as new data are collected. Future releases will include adjustments for gender and education level, subject to data availability. Patient-level data are not publicly available due to privacy restrictions.

#### Code availability

The OCTAL platform is the subject of intellectual property licensed by Oxford University Innovation, with Dr Sijia Zhao and Prof Masud Husain listed as inventors. Demonstrations of all OCTAL tasks are available at <http://octalportal.com>. The experimental and analysis code for the OCTAL tasks is distributed under a free academic licence administered by Oxford University Innovation. Interested researchers may apply for access via the OCTAL website (<https://octalportal.com/how-to-get-license-to-use-miniocal/>).

Received: 25 June 2025; Accepted: 6 January 2026;

Published online: 15 January 2026

#### References

- Nichols, E. et al. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. *Lancet Public Health* **7**, e105–e125 (2022).
- Alzheimer's Society. *Improving Access to a Timely and Accurate Diagnosis of Dementia in England, Wales and Northern Ireland*. [https://www.alzheimers.org.uk/sites/default/files/2023-05/alzheimers-society-improving-access-timely-accurate-diagnosis-england-wales-full-report\\_0.pdf](https://www.alzheimers.org.uk/sites/default/files/2023-05/alzheimers-society-improving-access-timely-accurate-diagnosis-england-wales-full-report_0.pdf) (2023).
- Nicholls, D. et al. *National Audit of Dementia: Spotlight Audit in Memory Assessment Services 2023/2024* (2024).
- 2025 Alzheimer's disease facts and figures. *Alzheimers Dement.* **21**, e70235 <https://doi.org/10.1002/alz.70235> (2025).
- Aliberti, M. J. R., Avelino-Silva, T. J. & Suemoto, C. K. Maximizing early dementia detection through medicare annual wellness visits. *JAMA Netw. Open* **7**, e2437162 (2024).
- Henrich, J., Heine, S. J. & Norenzayan, A. The weirdest people in the world?. *Behav. Brain Sci.* **33**, 61–83 (2010). discussion 83–135.
- Folstein, M. F., Folstein, S. E. & McHugh, P. R. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* **12**, 189–198 (1975).
- Nasreddine, Z. S. et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* **53**, 695–699 (2005).
- Hsieh, S., Schubert, S., Hoon, C., Mioshi, E. & Hodges, J. R. Validation of the Addenbrooke's cognitive examination III in frontotemporal dementia and Alzheimer's disease. *Dement. Geriatr. Cogn. Disord.* **36**, 242–250 (2013).
- Öhman, F., Hassenstab, J., Berron, D., Schöll, M. & Papp, K. V. Current advances in digital cognitive assessment for preclinical

- Alzheimer's disease. *Alzheimers Dement. Diagn. Assess. Dis. Monit* **13**, e12217 (2021).
11. Jutten, R. J. et al. Why a clinical trial is as good as its outcome measure: a framework for the selection and use of cognitive outcome measures for clinical trials of Alzheimer's disease. *Alzheimers Dement. J. Alzheimers Assoc.* **19**, 708–720 (2023).
  12. Soldan, A., Pettigrew, C. & Albert, M. Evaluating cognitive reserve through the prism of preclinical Alzheimer disease. *Psychiatr. Clin. N. Am.* **41**, 65–77 (2018).
  13. Huber, H. et al. A finger prick collection method for detecting blood biomarkers of neurodegeneration – a pilot study (DROP-AD). *Alzheimers Dement.* **19**, e080275 (2023).
  14. Polk, S. E. et al. A scoping review of remote and unsupervised digital cognitive assessments in preclinical Alzheimer's disease. *npj Digit. Med.* **8**, 1–14 (2025).
  15. Berron, D. et al. Remote and unsupervised digital memory assessments can reliably detect cognitive impairment in Alzheimer's disease. *Alzheimers Dement.* **20**, 4775–4791 (2024).
  16. Zhao, S. et al. Long COVID is associated with severe cognitive slowing: a multicentre cross-sectional study. *eClinicalMedicine* **68**, 102434 (2024).
  17. Zhao, S. et al. Rapid vigilance and episodic memory decrements in COVID-19 survivors. *Brain Commun.* **4**, fcab295 (2022).
  18. Spiers, H. J., Coutrot, A. & Hornberger, M. Explaining world-wide variation in navigation ability from millions of people: citizen science Project Sea Hero Quest. *Top. Cogn. Sci.* **15**, 120–138 (2023).
  19. Hampshire, A. et al. Cognitive deficits in people who have recovered from COVID-19. *eClinicalMedicine* **39**, 101044 (2021).
  20. Beishon, L. C. et al. Addenbrooke's Cognitive Examination III (ACE-III) and mini-ACE for the detection of dementia and mild cognitive impairment. *Cochrane Database Syst. Rev.* **12**, CD013282 (2019).
  21. McGraw, K. O. & Wong, S. P. Forming inferences about some intraclass correlation coefficients. *Psychol. Methods* **1**, 30–46 (1996).
  22. Koo, T. K. & Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **15**, 155–163 (2016).
  23. Fray, P. J., Robbins, T. W. & Sahakian, B. J. Neuropsychiatric applications of CANTAB. *Int. J. Geriatr. Psychiatry* **11**, 329–336 (1996).
  24. Falletti, M. G., Maruff, P., Collie, A. & Darby, D. G. Practice effects associated with the repeated assessment of cognitive function using the CogState battery at 10-minute, one week and one month test-retest intervals. *J. Clin. Exp. Neuropsychol.* **28**, 1095–1112 (2006).
  25. Pratap, A. et al. Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. *npj Digit. Med.* **3**, 1–10 (2020).
  26. Bridges, D., Pitiot, A., MacAskill, M. R. & Peirce, J. W. The timing mega-study: comparing a range of experiment generators, both lab-based and online. *PeerJ* **8**, e9414 (2020).
  27. Rodd, J. M. Moving experimental psychology online: How to obtain high quality data when we can't see our participants. *J. Mem. Lang.* **134**, 104472 (2024).
  28. Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N. & Evershed, J. K. Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behav. Res. Methods* **53**, 1407–1425 (2021).
  29. Salthouse, T. A. Selective review of cognitive aging. *J. Int. Neuropsychol. Soc.* **16**, 754–760 (2010).
  30. Seo, E. H. et al. A normative study of the Trail Making Test in Korean elders. *Int. J. Geriatr. Psychiatry* **21**, 844–852 (2006).
  31. Nielsen, T. R. Cognitive Assessment in culturally, linguistically, and educationally diverse older populations in Europe. *Am. J. Alzheimers Dis. Dementias* **37**, 1533317522117006 (2022).
  32. Briceño, E. M. et al. A cultural neuropsychological approach to harmonization of cognitive data across culturally and linguistically diverse older adult populations. *Neuropsychology* **37**, 247–257 (2023).
  33. Franzen, S. et al. A systematic review of neuropsychological tests for the assessment of dementia in non-Western, low-educated or illiterate populations. *J. Int. Neuropsychol. Soc.* **26**, 331–351 (2020).
  34. Toniolo, S. et al. Relationship of plasma biomarkers to digital cognitive tests in Alzheimer's disease. *Alzheimers Dement. Diagn. Assess. Dis. Monit.* **16**, e12590 (2024).
  35. Peirce, J. et al. PsychoPy2: experiments in behavior made easy. *Behav. Res. Methods* **51**, 195–203 (2019).
  36. Rey, A. L'examen psychologique dans les cas d'encephalopathie traumatique. *Arch. Psychol.* **28**, 286–340 (1941).
  37. Osterrieth, P. A. Le test de copie d'une figure complexe; contribution à l'étude de la perception et de la mémoire. *Arch. Psychol.* **30**, 206–356 (1944).
  38. Bowie, C. R. & Harvey, P. D. Administration and interpretation of the Trail Making Test. *Nat. Protoc.* **1**, 2277–2281 (2006).
  39. Sherman, E., Tan, J. & Hrabok, and M. A *Compendium of Neuropsychological Tests: Fundamentals of Neuropsychological Assessment and Test Reviews for Clinical Practice* (Oxford University Press, Oxford, New York, 2023).
  40. Zeng, Z., Miao, C., Leung, C. & Shen, Z. Computerizing Trail Making Test for long-term cognitive self-assessment. *Int. J. Crowd Sci.* **1**, 83–99 (2017).
  41. Jaeger, J. Digit symbol substitution test. *J. Clin. Psychopharmacol.* **38**, 513–519 (2018).
  42. Corsi, P. M. *Human Memory and the Medial Temporal Region of the Brain* (McGill University, Montreal, 1972).
  43. Pertzov, Y. et al. Binding deficits in memory following medial temporal lobe damage in patients with voltage-gated potassium channel complex antibody-associated limbic encephalitis. *Brain* **136**, 2474–2485 (2013).
  44. Zokaei, N. et al. Different patterns of short-term memory deficit in Alzheimer's disease, Parkinson's disease and subjective cognitive impairment. *Cortex* **132**, 41–50 (2020).
  45. Toniolo, S. et al. Performance and validation of a digital memory test across the Alzheimer's disease continuum. *Brain Commun.* **7**, fcab024 (2025).
  46. Zokaei, N. et al. Dissociable effects of the apolipoprotein-E (APOE) gene on short- and long-term memories. *Neurobiol. Aging* **73**, 115–122 (2019).
  47. Brady, T. F., Konkle, T., Alvarez, G. A. & Oliva, A. Visual long-term memory has a massive storage capacity for object details. *Proc. Natl. Acad. Sci. USA* **105**, 14325–14329 (2008).
  48. McKhann, G. M. et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement. J. Alzheimers Assoc.* **7**, 263–269 (2011).
  49. Petersen, R. C. et al. Mild cognitive impairment: clinical characterization and outcome. *Arch. Neurol.* **56**, 303–308 (1999).
  50. Jessen, F. et al. The characterisation of subjective cognitive decline. *Lancet Neurol.* **19**, 271–278 (2020).
  51. JASP Team. *JASP* (University of Amsterdam, 2022).
  52. Dunn, O. J. & Clark, V. Correlation coefficients measured on the same individuals. *J. Am. Stat. Assoc.* **64**, 366–377 (1969).
  53. Hittner, J. B., May, K. & Silver, N. C. A Monte Carlo evaluation of tests for comparing dependent correlations. *J. Gen. Psychol.* **130**, 149–168 (2003).
  54. Diedenhofen, B. & Musch, J. cocor: a comprehensive solution for the statistical comparison of correlations. *PLoS ONE* **10**, e0121945 (2015).
  55. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
  56. Salarian, A. Intra-class Correlation Coefficient (ICC). MATLAB Central File Exchange. <https://uk.mathworks.com/matlabcentral/fileexchange/22099-intra-class-correlation-coefficient-icc>. Retrieved January 11, 2026.

## Acknowledgements

S.Z., S.T., A.S., A.G., M.J.B. and M.H. were funded by the Wellcome Trust (226645/Z/22/Z). S.T., S.G.M. and M.H. were supported by National Institute of Health and Care Research (NIHR) Oxford Health Biomedical Research Centre (BRC). S.G.M. was funded by a Medical Research Council (MRC) Clinician Scientist Fellowship (MR/P00878/X), NIHR Oxford BRC. Q.-Y. T. was funded by the Guangdong-Hong Kong Universities "1+1+1" Joint Funding Programme. C.G. received a residency scholarship, funded by the Italian Ministry of University and Research (MUR), as part of the Neurology Residency Program at the Univeristà degli Studi di Milano. We gratefully acknowledge all participants and their families for their time and commitment, as well as the staff of the Cognitive Disorders Clinic at the John Radcliffe Hospital for their support. We particularly thank those participants who provided valuable feedback during the design and piloting of OCTAL, which helped improve the platform's usability. We also thank Dr Gabriel Jones for his contribution to the branding concept for OCTAL.

## Author contributions

S.Z., S.To., S.M., and M.H. conceived and designed the study. S.Z. developed and maintained the OCTAL platform. S.Z., S.To., A.S., A.G., C.G., M.J.B., S.M., S.Th., and M.H. were responsible for participant recruitment and data collection. S.Z., S.To., A.S., A.G., and M.J.B. conducted the quality assurance of the remote data. S.Z. curated and analysed the data. Q.Y.T. also contributed to statistical analysis. S.Z. prepared all figures. S.Z. and M.H. prepared the original manuscript draft. All authors reviewed and edited the paper.

## Competing interests

Sijia Zhao and Masud Husain are named inventors on the OCTAL intellectual property, which is licensed by Oxford University Innovation. All other authors declare no competing financial or non-financial interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41746-026-02346-6>.

**Correspondence** and requests for materials should be addressed to Sijia Zhao.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026