Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Beyond the Clinic: Bridging the Gap in Post-Stroke Cognitive Monitoring with the Oldenburg Test Battery for Remote Digital Phenotyping of Post-Stroke Cognition

Paul Leonard Grupe
paul.grupe@uol.de

Carl von Ossietzky University of Oldenburg https://orcid.org/0009-0005-6581-5280

Helmut Hildebrandt
Carl von Ossietzky University of Oldenburg

Karsten Witt
Carl von Ossietzky University of Oldenburg

Andreas Kastrup
Klinikum Bremen-Mitte

Christiane Margarethe Thiel
Carl von Ossietzky University of Oldenburg

Mandy Roheger
Carl von Ossietzky University of Oldenburg

Andrea Hildebrandt
andrea.hildebrandt@uni-oldenburg.de

Carl von Ossietzky University of Oldenburg

Research Article

Keywords:

Posted Date: September 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11009571/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Cognitive diagnosis post-stroke is typically restricted to static, in-clinic assessments. We developed and evaluated the *Oldenburg Test Battery for Remote Digital Phenotyping of Post-Stroke Cognition (ORPheoS)*, which is an item bank and test battery to be used for sampling brief test versions for smartphone-based repeated testing in ecologically valid longitudinal cognitive phenotyping studies to inform stroke rehabilitation.

A total of 280 controls and 232 stroke patients underwent ORPheoS during one baseline measurement at ~five days post-stroke, comprising five tasks targeting reasoning, recognition and associative memory, and external and internal interference control. A planned-missingness design with linkage items enabled joint calibration across partially overlapping item subsets. Item properties and domain specific ability scores were estimated using Rasch-family item response theory-based latent trait uni- and multidimensional models, as well as mixed-effects modelling. A bifactor S-1 model was used to examine the taxonomic order of the scaled abilities measured by ORPheoS.

Latent trait models demonstrated acceptable to excellent item properties. Domain specific split-half reliability estimates ranged from 0.52–0.88. Known-groups validity was domain-specific. Reasoning showed a large group difference ($g = 0.45$, $p < 0.001$) and the strongest associations with global cognition. Associative memory and mental speed also differentiated groups ($g = 0.25$, $p = 0.011$; and $g = 0.91$, $p < 0.001$). Familiarity-based recognition showed no group difference ($g = -0.03$, $p = 0.734$), while stroke participants outperformed controls on Recollection-based discrimination ($g = -0.48$, $p < 0.001$). The bifactor S-1 model showed acceptable fit (scaled $\chi^2(47) = 161.12$, $p < 0.001$; robust CFI = 0.941; robust RMSEA = 0.078), confirming interference control and memory as domain specific factors beyond general cognitive ability indicated by reasoning.

ORPheoS provides psychometrically calibrated, domain-specific tasks and items within a scalable digital framework, establishing a measurement foundation for longitudinal and personalised monitoring of cognitive functions after stroke. The current dataset and estimated item properties allow for adaptive sampling and item generation for repeated ecological assessments of cognition after stroke, with potential of supporting tailor-made rehabilitation planning and monitoring.

This work was supported by the Research Training Group (RTG) 2783, funded by the German Research Foundation (DFG) – Project ID 456732630.

1. Introduction

Stroke is a primary cause of global long-term disability [1]. Although rehabilitation research has historically prioritised motor recovery, cognitive dysfunction remains a critical barrier to autonomy, predicting lower quality of life [2] and slower motor improvement [3]. Recent advances in wearable technologies have transformed motor rehabilitation monitoring [4–7], yet comparable cognitive innovations are just emerging

[8–10], with open tools completely absent. To enable individualised, effective rehabilitation, cognition must be monitored with the precision and *in-situ* focus now common in the motor domain.

Post-stroke cognitive performance varies across days and contexts [11,12]. Developing personalised, dynamically adaptable interventions therefore requires measurement tools capable of capturing within-person fluctuations and change over time. Neuropsychological assessments typically offer highly controlled diagnostically relevant static snapshots of ability [11,13]. The dynamic nature of cognition in real-world environments however is not reflected in those snapshots. To guide adaptive interventions and track meaningful change, assessments must be sensitive to momentary variations, feasible for remote repeated administration, and resilient to practice effects.

Ecological momentary assessment (EMA) provides a powerful framework for dynamic in-situ measurement [14]. However, stroke-related EMA research largely relies on self-report data [15], which is prone to bias [16] and lacks objective performance metrics. Applying EMA principles to post-stroke cognitive assessment poses unique challenges. EMA requires brief, engaging, smartphone-based and adaptive applications of cognitive tasks and systematic sampling of items to avoid repetition of content (stimuli) and mitigate the effects of practice. Yet, ensuring comparability across testing repetitions requires items with equivalent discrimination power to inform sampling. No multi-domain item bank with homogeneous discrimination currently exists for post-stroke cognitive measurement.

Recent developments, such as the IC3 battery [8] and the Oxford Cognitive Testing Portal [17] demonstrate the feasibility of remote, digital cognitive testing in stroke and aging/dementia populations respectively. However, robust longitudinal measurement requires psychometrically calibrated item banks with equivalent item properties. Equivalence can be tested via Rasch-family Item Response Theory (IRT) models [18–20]. The fundamental advantage of Rasch-equivalent items is specific objectivity [19]: an individual's latent trait estimate remains independent of the specific item subset administered. Because Rasch-homogeneous items are uniform in discrimination power and differ only by difficulty, they are psychometrically interchangeable [19–21]. This allows comparable trait estimation across testing sessions within and between participants, even when item content changes.

We developed the *Oldenburg Test Battery for Remote Digital Phenotyping of Post-Stroke Cognition (ORPheoS)*, a brief, repeatable, smartphone-based cognitive battery designed for administration in daily life, aligning post-stroke assessment with the modernisation of measurement through item banks and latent trait modelling [20,22,23]. Complementing broader tools like IC3, ORPheoS restricts domain coverage to mental speed, executive functioning, memory, and general cognition, using a nearly language-free design to isolate cognitive phenotypes from aphasic confounds. All tasks were designed to conform to Rasch-family IRT models, formally establishing an item bank that supports adaptive sampling for future EMA studies. This paper evaluates psychometric properties, reliability, and construct validity of ORPheoS, and discusses its potential for longitudinal cognitive recovery monitoring.

2. Methods

2.1 Participants

After data cleaning, the sample comprised 232 early sub-acute stroke survivors and 280 matched community controls (Supplementary Table 1). Controls were recruited in two phases (Feb–Nov 2023; Mar–Apr 2026); the stroke cohort was recruited continuously (Aug 2024–Apr 2026) from inpatient units at the Evangelisches Krankenhaus Oldenburg and Klinikum Bremen-Mitte. Stroke inclusion criteria were a confirmed diagnosis, German proficiency, and sufficient visual-motor ability to navigate a touchscreen. Patients predominantly had minor strokes (*Mdn* NIHSS = 2) preparing to transition to independent living. Controls with a history of stroke, TIA, or dementia were excluded. The stroke sample (M_{age} = 64 years, SD_{age} = 13.29) was tested 5.46 days (*SD* = 3.08) post-stroke; the control sample mean age was 62 years (*SD* = 8.27). Following written informed consent, the study adhered to the Declaration of Helsinki (excluding preregistration) and was approved by the Medical Ethics Committee of the Carl von Ossietzky Universität Oldenburg (Ref. Nos. 2022-093, 2024-033).

2.2 Procedure

Analyses used baseline data from an ongoing longitudinal study. GDPR-compliant data were collected via the SoSci-Survey platform [24] on secure servers. Stroke participants were tested in-person at the clinics using university-owned smartphones (Motorola Moto G). To ensure feasibility, a planned-missingness design was deployed: the initial 50 patients completed larger task subsets, whereas subsequent participants received non-overlapping item blocks linked by shared 'anchor' items to enable joint calibration onto a common latent metric. Controls completed the battery remotely and unsupervised on their own devices over two to five consecutive days. Despite setting differences (supervised vs remote), stimuli, instructions, practice trials, and task order were identical across groups. Controls received monetary remuneration. Except for the pen-and-paper Montreal Cognitive Assessment (patients only, MoCA; [25]), all tasks were administered digitally.

2.3.1 The ORPheoS Battery

ORPheoS provides a scalable digital phenotyping framework targeting cognitive domains that are clinically consequential post-stroke and psychometrically amenable to repeated assessment: Interference control (external and internal), mental speed, reasoning, and memory (Fig. 1; 24,25). Domain selection prioritised executive and mnemonic functions that constrain goal-directed behaviour; domains like complex aphasia or social cognition were excluded to maintain compatibility with unidimensional latent trait modelling. Intuitive touchscreen interfaces and standardised visual stimuli minimised confounding variance from language proficiency and fine-motor coordination, though baseline visual and motor capacities remain prerequisites for administration. Items share uniform discrimination power and vary only along a single dimension of difficulty, rendering them psychometrically interchangeable. This EMA-compatible structure permits administering non-overlapping item subsets across sessions,

facilitating comparable trait score estimation without repeated stimulus exposure. **Figure 1: Theoretical taxonomic order of the ORPheoS task battery.** Targeted domains are divided into executive function (left), memory (right), and general cognitive ability. Executive function measurement includes the SART-ED (Sustained Attention to Response Task with External Distraction; measures external interference control, based on the Go-No-Go paradigm) and the CI task (Cognitive Interference task; measures internal interference control; based on A-B/A-C/A-D paradigm). Reasoning (operationalized by the Geometric Reasoning Task, GRT) and mental speed (operationalised by the SART-ED "Go" trials without a distractor) serve as markers for general cognitive functioning. Memory comprises the DRT (Delayed Recognition Task of objects; measures recognition after incidental encoding) and the PAL (Paired Association Learning task; measures associative memory). The path diagram reflects a bifactor S-1 [28] type structural representation, illustrating both digital interfaces and the underlying taxonomy of the measured abilities.

Sustained Attention to Response Task with External Distraction (SART-ED): A fixed-paced Go-No-Go paradigm comprising 200 rapid serial digits (1–9) presented for 500 ms, with '3' as the No-Go stimulus (60 trials). Distractors (neutral faces above or below digits) appear in 60 random trials (30 Go, 30 No-Go). Responses require a single screen tap. Two performance indicators are derived: mental speed (Go-trial reaction times) and external interference control (resilience to distraction, isolated via residualisation; see below). To account for trial-level temporal dependencies violating standard IRT assumptions, external interference control is modelled via a two-step mixed-effects pipeline (Section 2.5).

Cognitive Interference (CI) task

Nine five-item testlets utilising semantically related (A–B), unrelated (A–C), and novel (A–D) word-pairs. The 'A' word is presented, and the target (B, C, or D) must be recognised from a distractor list. Lures consisting of previously paired words induce interference. Baseline recall (A–C) and recall under interference (A–D) are modelled separately to isolate internal interference control from general memory. Excluding this task renders ORPheoS fully language-free.

Geometric Reasoning Task (GRT)

59 pattern-matching items presenting a vertical sequence of eight symbols. Participants identify the symbol differing in size/shape or violating a geometric progression rule ('Which symbol does not fit in?'; [29]. Scoring is binary (1 = correct, 0 = incorrect).

Delayed Recognition Task (DRT)

40 object-recognition trials administered approximately five minutes after incidental encoding during a prior visual matching task (not analysed here due to design-inherent ceiling effects). Participants identify the target among two distractors and an "object unknown" option. To also measure recollection-based

discrimination beyond familiarity-based recognition, 18 trials contain only novel stimuli, requiring the "object unknown" response.

Paired Association Learning (PAL) task

An adaptive associative memory task where objects presented in specific locations must be recalled in an altered order. To prevent item-specific learning across adaptive difficulty levels, the PAL utilises four parallel versions. Difficulty increases across five levels (2, 3, 4, 6, and 8 pairs), with up to three attempts per level. Performance is quantified as percentage correct per level.

Table 1
summarises task specifications across the ORPheoS battery, including measured abilities, item pool sizes, planned-missingness item allocations per patient subgroup, and response formats.

Task	Measured ability	Origin	Item pool (*K*)	Items (*K*) per patient subgroup (*n* ≈ 50)	Response format
SART-ED	*Resilience to external interference*	Go-No-Go paradigm	200	200	Single-tap (go)
	Mental speed				
CI	*Resilience to internal interference*	AB-AC-AD paradigm	nine testlets / five items each	one testlet	Multi-choice
GRT	*Reasoning*	Geometric Reasoning Task inspired by LPS-2 [29]	59	10	8-alternative-forced-choice
DRT	*Familiarity-based recognition*	Figural delayed recognition paradigms	40	20	4-alternative-forced-choice
	Recollection-based discrimination				
PAL	*Associative memory*	Figural association memory paradigms	four versions / five levels each	one version	8-alternative-forced-choice

Table 1: **Specifications of the ORPheoS battery.** Abbreviations: SART-ED, Sustained Attention to Response Task with External Distraction; CI, Cognitive Interference; GRT, Geometric Reasoning Task; DRT, Delayed Recognition Task; PAL, Paired Association Learning task; LPS-2, Leistungsprüfsystem Version 2.

Item distribution (*K*) follows a planned missingness design to reduce participant burden. SART-ED was administered with all items to facilitate reliability of reaction time metrics. Stimuli for GRT are inspired from the LPS-2 [29], while DRT and PAL utilise standardised figural paradigms with familiarity-based recognition lures in the DRT to control for false-alarm rates. All tasks are delivered via a touchscreen interface using standardised grid layouts to ensure uniform motor requirements across the 4- and 8-alternative forced-choice formats.

2.3.2 Montreal Cognitive Assessment (MoCA)

The MoCA [25], an established paper-and-pencil measure of global cognition (attention, memory, language, executive function, visuospatial ability), facilitated convergent and divergent validation. Scores were converted to percentage-correct. As controls did not complete the MoCA, validity estimations were restricted to the stroke sample.

2.4 Data Preprocessing

Raw data were processed in R (4.4.3). Not administered tasks and skipped items were treated as missing (NA) without imputation. Completion rates ranged from 65.52% to 97.41% for stroke participants and 70.36% to 100% for controls. This variance reflected the two-phase control recruitment (Phase 1: DRT and GRT only; Phase 2: full battery) and patient attrition during the 90-minute session. Domain-specific latent ability parameters (θ-scores) were derived from all available responses using full-information maximum likelihood (FIML) estimation within an item response theory (IRT) framework. No further participant exclusions were applied.

2.5 Psychometric Analyses

Item analysis and scaling of domain-specific abilities

Dimensionality was assessed via Principal Component Analysis of tetrachoric correlation matrices. Items with substantial secondary loadings or model misfit (Infit/Outfit Mean-Square [MNSQ] outside 0.5–1.5) were iteratively removed if review indicated ability-irrelevant variance. Domain-specific responses were modelled using unidimensional Rasch (1-parameter logistic; 1PL) models, evaluated against 2-parameter logistic (2PL) alternatives via AIC, BIC, and likelihood-ratio tests. A multidimensional IRT approach was adopted for the DRT following evidence of a two-dimensional structure.

To isolate distraction effects from baseline performance in the SART-ED trial-level data, we used a two-step mixed-effects pipeline for ability scaling. First, linear and generalised linear (binomial, logit link) mixed-effects models predicted log-transformed reaction times (correct Go-trials) and Go-trial accuracy, respectively. Both models regressed performance on trial condition (distraction vs baseline), mean-centred baseline performance, and their interaction, specifying random intercepts and slopes for condition. Second, participant-specific Empirical Bayes random slopes (representing the condition effect adjusted for baseline) were extracted and sign-aligned so that higher values indicated greater resilience

(RT slopes multiplied by -1). Finally, Minimum Residual Factor Analysis derived a single θ-score for external interference control.

Scale reliability estimation

Split-half reliability was estimated for each cognitive domain using parallel half-task parcels (a vs. b). We used odd-even splits for GRT, DRT, and CI, and an interleaved split (levels 1, 3, 5 vs. 2, 4) for the progressive difficulty of the PAL. For SART-ED, Empirical Bayes random slopes were estimated separately for odd and even trials. Except for SART-ED, parcel θ-scores were derived via Expected A Posteriori (EAP) scoring from independent 1PL models. Reliability was calculated as the Spearman-Brown corrected Pearson correlation between parcel-specific θ-scores.

Validity estimation

Final θ-scores were estimated using marginal maximum likelihood with EAP scoring [30]. Construct validity was assessed through inter-task associations, correlations with MoCA scores (stroke sample only), and a bifactor S-1 confirmatory factor analysis model. Criterion validity was examined via known-groups comparisons between stroke survivors and controls.

Role of the funding source

This study was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project-ID 456732630 - RTG 2783 "Neuromodulation of Motor and Cognitive Function in Brain Health and Disease". The funding source had no role in the study design, in the data collection, analysis, and interpretation of data, in the writing of the report, and in the decision to submit the paper for publication.

3. Results

3.1 Psychometric Properties of the ORPheoS Tasks

Dimensionality analysis and item selection

Combined dimensionality analysis on the stroke and control sample supported unidimensional structures for GRT and PAL, and a two-dimensional structure for DRT. Item Principal Component Analysis of the 59 GRT items revealed a dominant first component (36.1% variance explained; eigenvalue ratio 6.23); excluding items with secondary loadings (k = 4) or heterogeneous discrimination (k = 6), yielded a final Rasch-calibrated set of 49 well-fitting items. For the DRT, Principal Component Analysis of 40 items supported a two-dimensional structure (eigenvalues 14.28 and 6.63): Familiarity-based recognition (k = 21) and recollection-based discrimination (k = 11). The PAL was best represented by a unidimensional partial-credit Rasch model, outperforming a graded response alternative (ΔAIC = 105.1, $\chi^2(4) = 97.08$, $p < 0.001$). These results confirm that the item bank exhibits uniform item discrimination and specific objectivity, rendering it suitable for adaptive item sampling in EMA studies.

Before residualisation, CI baseline recall (A-C) and interference recall (A-D) were strongly correlated (r = 0.79). Residualisation successfully isolated an internal interference score orthogonal to baseline ($r \approx 0$) yet correlated with raw interference performance (r = 0.61). SART-ED mixed-effects models confirmed a robust distraction effect: Increased Go-trial reaction times (β = 0.114, *SE* 0.004; *Wald t* = 26.7) and reduced Go accuracy (β = −0.152, *SE* 0.070; *Wald z* = − 2.19, p = 0.029).

Reliability

Split-half reliability estimates (Fig. 3, diagonal) varied across domains, reflecting distinct scaling challenges. Traditional ability estimates demonstrated high reliability: PAL and GRT ($r_{x'x''}$ = 0.86), DRT ($r_{x'x''}$ = 0.88) and SART-ED ($r_{x'x''}$ = 0.93). Conversely, interference composites yielded lower but acceptable reliabilities: SART-ED ($r_{x'x''}$ = 0.62) and CI ($r_{x'x''}$ = 0.52). These lower estimates reflect the inherent challenge of residualised or difference-score executive metrics: removing reliable baseline variance that correlates highly with the alternative condition score restricts the residual variance available to represent the interference-specific effect.

Validity

Known-groups validity. Latent ability scores (θ; Fig. 2) revealed no significant group differences for external (Hedges' g = 0.16, 95% *CI* − 0.06 to 0.37, p = 0.142) or internal interference (g = − 0.13, 95% *CI* − 0.35 to 0.10, p = 0.309). Controls significantly outperformed stroke survivors in mental speed (g = 0.91, 95% CI [0.68, 1.14], p < 0.001), reasoning (g = 0.45, *CI* 0.27–0.63, p < 0.001) and associative memory (g = 0.25, *CI* 0.06–0.45, p = 0.011). Familiarity-based recognition (DRT D1) showed no group difference (g = − 0.03, *CI* − 0.22 to 0.16, p = 0.734) whereas stroke participants outperformed controls on recollection-based discrimination (g = − 0.48, *CI* − 0.67 to − 0.28, p < 0.001). These differences remained significant after controlling for education (Supplementary Table 5) and sex (Supplementary Table 6), confirming the results are not driven by sample imbalances in demographics.

Convergent and divergent validity. Inter-task correlations (Fig. 3) supported convergent validity among core abilities: Reasoning, familiarity-based recognition, associative memory, and global cognition (MoCA) showed moderate positive associations ($r \approx 0.25–0.48$). Mental speed exhibited a distinct profile, correlating positively with reasoning (r = 0.25) and associative memory (r = 0.16), but showing no significant association with familiarity-based recognition (r = -0.07). Recollection-based discrimination showed partial discriminant validity, correlating with global cognition (r = 0.34) and familiarity-based recognition (r = 0.35), but only weakly with reasoning (r = 0.07) and associative memory (r = 0.10) and inversely with mental speed (r = -0.20). Residualised interference scores correlated negligibly with most other measures ($rs \leq .10$), except for a weak association between residualised SART-ED and mental speed (r = 0.17). This is consistent with their construction as capacity-independent sensitivity indices. Comparisons with non-residualised performance indices (Supplementary Table 4) confirm that processing speed drives overall inter-task associations and group differentiation. Association patterns varied slightly between samples, potentially reflecting differences in testing supervision.

Taxonomic model of abilities measured by the ORPheoS battery

An S-1 bifactor model evaluating the theoretical structure (Fig. 4) demonstrated acceptable fit (scaled $\chi^2(47) = 161.12$, $p < 0.001$; robust CFI = 0.941; robust TLI = 0.917; robust RMSEA = 0.078, 90% CI 0.064–0.093; SRMR = 0.088). The general cognitive ability factor (*g*) accounted for substantial variance across domain-specific θ-scores, loading most strongly on reasoning indicators (0.77–0.96). Beyond *g*, external (~ 0.67) and internal interference control (~ 0.60) emerged as robust, distinct factors. Notably, except for one, external interference indicators did not load significantly onto *g*. This suggests that resilience to distraction is a cognitive domain largely orthogonal to global ability, though this loading pattern additionally reflects scaling challenges of residualised executive control metrics. Associative memory and recollection-based recognition formed specific factors moderately associated with *g*, consistent with their intermediate taxonomic order. Covariance between these latent factors was freely estimated to account for local dependence within the hierarchical task structure. **Figure 4: S-1 Bifactor model of cognitive abilities as measured by ORPheoS.** The path diagram illustrates the latent structure of parcel-level θ-scores. The general cognitive factor (*g*, top) accounts for shared variance across all tasks, reflecting global cognitive ability, with geometric reasoning acting as the reference domain. Five specific factors (bottom) capture additional domain-specific variance, independent of *g*. Directed paths represent standardised factor loadings; rectangles represent θ-scores estimated from half-task item parcels (odd/even or interleaved). A dashed bidirectional arc indicates a covariance between the latent associative memory and familiarity-based recognition specific factors, accounting for local dependence per design. Additional residual covariances between corresponding theta scores of the EI and MS scores (parcel a with parcel a; parcel b with parcel b) were included to account for task-specific shared method variance. EIC: External Interference Control as derived from the SART-ED task; IIC: Internal Interference Control as derived from the CI task; MS: Mental Speed as derived from the SART-ED task; R: Reasoning as derived from the GRT; Rec: Recollection-based Recognition as derived from the DRT (without recollection-based discrimination items); AM: Associative Memory as derived from the PAL; a vs b is indicating the odd/even item parcels used for theta estimation.

4. Discussion

4.1 Main Findings

This study establishes the psychometric foundation, a calibrated digital measurement framework and validated item bank, required prior to deploying high-frequency ecological momentary assessment (EMA) in post-stroke populations. Such a framework is a prerequisite for personalised and intense long-term monitoring of cognitive rehabilitation after stroke, a task that traditional, static neuropsychological assessments are not suitable to perform. Using a planned-missingness design, we demonstrated that robust latent trait estimation is achievable despite only partial individual-level data. Reasoning and associative memory emerged as structurally robust, unidimensional ability estimates. In contrast, recognition differentiated into two dimensions: recognition of previously presented stimuli, driven by familiarity and recollection-based discrimination (indicated by correct rejections) of novel stimuli.

Furthermore, modelling condition differences successfully isolated interference control from general cognitive ability. Notably, mental speed, reasoning, and associative memory demonstrated significant group differences, whereas recognition ability did not. This indicates the battery functions not merely as a homogeneous impairment screener, but exhibits domain-specific discriminative sensitivity. Reasoning emerged as the most clinically discriminative domain, capturing higher-order impairments even within this predominantly minor-stroke cohort. Future research must determine how these tasks differentiate between minor and severe stroke profiles.

A central finding is that ORPheoS tasks measure executive functioning with differential sensitivity to stroke-related impairment. While residualised interference control showed minimal group differences, domain-general mental speed and reasoning revealed substantial deficits in the stroke cohort. Cross-sectional cognitive burden following minor stroke thus appears primarily driven by reductions in higher-order abilities rather than specific failures in inhibitory control. This pervasive reduction in mental speed and higher-order processing likely drives the observed qualitative shift in recognition related decision-making. Notably, stroke participants outperformed controls on recollection-based discrimination, suggesting a more cautious response criterion or "object unknown" default decision when confidence is low [31]. This cautiousness was specific to novel stimuli; post-hoc analysis confirmed no increased "object unknown" rate when targets were present. The battery captures both quantitative speed deficits and strategic cognitive adaptations, qualitative differences warranting further investigation in more severely impaired cohorts.

The cross-domain bifactor model confirms the utility of separating higher-order abilities from domain-specific interference control and memory. While mental speed and reasoning act as strong cross-sectional discriminators of impairment, isolating interference control and memory provides orthogonal metrics. Tracking intra-individual fluctuations in these domains, relative to higher-order abilities, is essential for future EMA; temporary lapses in interference control and response caution may specifically predict momentary fatigue or functional errors beyond general speed and reasoning. Consequently, ORPheoS complements static assessments and deconstructs cognitive variance for dynamic, high-frequency monitoring [11].

4.2 Methodological Implications for Clinical Use

This work demonstrates that rigorous psychometric modelling can be integrated into a scalable digital framework without compromising clinical feasibility. The nearly language-free design ensures measured deficits reflect higher-order cognitive abilities rather than primary sensory, motor, or aphasic impairments. A key strength is the foundation in a calibrated, Rasch-conform item bank. By ensuring items meet the requirement of specific objectivity [18], person ability estimates are mathematically independent of the specific item subset administered, provided item difficulty is matched. This psychometric constraint enables the transition from static to adaptive assessments.

Furthermore, stimulus properties identified here can guide automatic item generation. By deconstructing cognitive requirements, AI-based approaches can generate new items calibrated to the established

psychometric properties [22]. Our findings allow for increased precision in predicting item difficulty, streamlining the incorporation of novel content.

Finally, our results provide the empirical basis for strategic item sampling in EMA. By focusing on domains susceptible to impairment post-minor stroke, such as mental speed, reasoning, and spatial memory, adaptive item selection algorithms can be implemented to optimise ability estimates. Because items are Rasch-homogeneous, different subsets can be used daily without compromising score equivalence. This facilitates weighted, shortened EMA bursts, maximising monitoring efficiency within limited smartphone measurement windows. While the current iteration of ORPheoS captures a deliberately constrained breadth of functions to preserve its largely language-free- and minimal motor-demands, the underlying psychometric architecture is modular and additional domains can be added in the future. Similar to the HERMeS trial in heart failure [6], proactive monitoring of targeted higher-order functions could trigger timely therapeutic adaptations during sub-acute recovery. Procedural consistency, ensured by GDPR-compliant infrastructure and fixed stimuli, positions ORPheoS as a calibrated measurement system rather than a mere test battery.

4.3 Future Directions

Our psychometric findings rely on cross-sectional baseline data from an acute stroke cohort; sensitivity to change and within-person reliability across measurement occasions remain to be established. While the supervised (stroke) versus remote (control) testing asymmetry is a potential confound for known-groups validity, successful completion by controls demonstrates the feasibility of remote ORPheoS application.

The sample was entirely white/Caucasian and German-speaking, with a male-dominant clinical group. To move beyond "WEIRD" population biases and ensure universal item parameter invariance, future research must involve ethnically diverse populations, building on cross-cultural validation approaches already demonstrated for comparable remote digital batteries [17]. Finally, as participation required sufficient motor and cognitive ability to operate a smartphone, the current framework excludes severely impaired individuals. Future iterations should incorporate multimodal interfaces, such as voice-activated responses or simplified visual-motor requirements, to extend momentary cognitive monitoring to the full spectrum of stroke recovery.

5. Conclusion

ORPheoS provides a psychometrically calibrated infrastructure for objective cognitive assessment in post-stroke rehabilitation. Through latent trait modelling and an S-1 bifactor structure, we established that the battery yields reliable, domain-specific estimates, with reasoning and mental speed emerging as the most robust discriminators between stroke and control individuals. Anchoring the battery in a Rasch-homogeneous item bank provides the infrastructure for algorithmic item sampling in EMA and a generative foundation for future AI-based expansion. ORPheoS thus offers a rigorous psychometric

basis to pursue the predictive value of digital cognitive phenotyping in longitudinal studies and EMA-based rehabilitation monitoring.

Declarations

Data sharing statement

All assessed test items as well as deidentified individual participant data and a data dictionary defining each field will be made available to researchers upon reasonable request to the corresponding authors. Access will be granted following the approval of a research proposal and the execution of a signed data access agreement. Furthermore, the full dataset and item bank is intended for public release through a dedicated data paper in the future. All analysis scripts used in this study are publicly available on GitHub at https://github.com/xxplgruxx/EMA4Stroke.

Author contributions

PLG, HH, CMT, MR and AH conceptualised the study and designed the ORPheoS battery. PLG, KW and AK were involved in participant recruitment and the conceptualisation of the data acquisition. PLG and AH developed the statistical analysis plan. PLG performed the data cleaning, psychometric modelling, and primary statistical analyses and AH supervised the process of analyses. PLG drafted the manuscript. All authors (PLG, HH, KW, AK, CMT, MR, AH) critically reviewed the manuscript for intellectual content and approved the final version for submission. PLG and AH directly accessed and verified the underlying analyses reported in the manuscript. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

References

1. El Husseini N et al (2023) Cognitive Impairment After Ischemic and Hemorrhagic Stroke: A Scientific Statement From the American Heart Association/American Stroke Association. Stroke 54:e272–e291
2. Stolwyk RJ et al (2024) Post-stroke Cognition is Associated with Stroke Survivor Quality of Life and Caregiver Outcomes: A Systematic Review and Meta-analysis. Neuropsychol Rev 34:1235–1264
3. Einstad MS et al (2021) Associations between post-stroke motor and cognitive function: a cross-sectional study. BMC Geriatr 21:103
4. Bui Q et al (2022) Ecological Momentary Assessment of Real-World Functional Behaviors in Individuals With Stroke: A Longitudinal Observational Study. Arch Phys Med Rehabil 103:1327–1337
5. Sun S, Simonsson O, McGarvey S, Torous J, Goldberg SB (2024) Mobile phone interventions to improve health outcomes among patients with chronic diseases: an umbrella review and evidence synthesis from 34 meta-analyses. Lancet Digit Health 6:e857–e870
6. Yun S et al (2025) Evaluation of mobile health technology combining telemonitoring and teleintervention versus usual care in vulnerable-phase heart failure management (HERMeS): a

multicentre, randomised controlled trial. Lancet Digit Health 7

7. Adans-Dester C et al (2020) Enabling precision rehabilitation interventions using wearable sensors and machine learning to track motor recovery. Npj Digit Med 3:121
8. Gruia D-C et al (2025) Development and Validation of the IC3: An Online Remote Assessment Technology for Deep Phenotyping and Monitoring of Cognitive Impairment After Stroke. Assessment 10731911251381572. 10.1177/10731911251381572
9. Shi Y et al (2026) Self-Administered Mobile Cognitive Assessments for Individuals with Stroke: A Feasibility and Validation Study. Arch Phys Med Rehabil S0003–9993(26):00690–00698. 10.1016/j.apmr.2026.04.030
10. Bălăeţ M et al (2024) Online cognitive monitoring technology for people with Parkinson's disease and REM sleep behavioural disorder. Npj Digit Med 7:118
11. Weizenbaum E, Torous J, Fulford D (2020) Cognition in Context: Understanding the Everyday Predictors of Cognitive Performance in a New Era of Measurement. JMIR MHealth UHealth 8:e14328
12. Hawks ZW et al (2024) Dynamic associations between glucose and ecological momentary cognition in Type 1 Diabetes. Npj Digit Med 7:59
13. Zdanowski S, Tieks A, Jeronimus BF, Zuidersma M (2022) Intra-Individual Variability in Cognitive Performance Can Befuddle the Study of Cognitive Impairments and Decline. J Alzheimer's Dis 85:519–525
14. Shiffman S, Stone AA, Hufford MR (2008) Ecological momentary assessment. Annu Rev Clin Psychol 4:1–32
15. Lau SCL, Bright L, Connor LT, Baum CM (2025) Experiences With Mobile Health-Enabled Ambulatory Monitoring Among Stroke Survivors: A Qualitative Study. OTJR Occup Particip Health 45:232–240
16. Spitzer S, Weber D (2019) Reporting biases in self-assessed physical and cognitive health status of older Europeans. PLoS ONE 14:e0223526
17. Zhao S et al (2026) Remote digital cognitive assessment for aging and dementia using the Oxford Cognitive Testing Portal OCTAL. Npj Digit Med 9:162
18. Tesio L (2003) Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. J Rehabil Med 35:105–115
19. Gibbons CJ (2017) Turning the Page on Pen-and-Paper Questionnaires: Combining Ecological Momentary Assessment and Computer Adaptive Testing to Transform Psychological Assessment in the 21st Century. Front Psychol 7
20. da Rocha NS, Chachamovich E, de Almeida Fleck MP, Tennant A (2013) An introduction to Rasch analysis for Psychiatric practice and research. J Psychiatr Res 47:141–148
21. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ (2007) Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. Lancet Neurol 6:1094–1105

22. Sommer M, Arendasy M (2025) Automatic- and Transformer-Based Automatic Item Generation: A Critical Review. J Intell 13:102
23. Capon W et al (2026) Validating an adaptive digital assessment of youth mental health needs: a cross-sectional study. Npj Digit Med 9:173
24. Leinert DJ (2024) SoSci Survey. SoSci Survey GmbH
25. Nasreddine ZS et al (2005) The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. J Am Geriatr Soc 53:695–699
26. Mancuso M et al (2023) The impact of cognitive function deficits and their recovery on functional outcome in subjects affected by ischemic subacute stroke: results from the Italian multicenter longitudinal study CogniReMo. Eur J Phys Rehabil Med 59:284–293
27. O'Sullivan MJ, Li X, Galligan D, Pendlebury ST (2023) Cogn Recovery After Stroke: Memory Stroke 54:44–54
28. Eid M, Geiser C, Koch T, Heene M (2017) Anomalous results in G-factor models: Explanations and alternatives. Psychol Methods 22:541–562
29. Kreuzpointner L, Lukesch H, Horn W (2013) Leistungsprüfsystem 2. LPS-2. Manual. Hogrefe, Göttingen
30. Bock RD, Mislevy RJ (1982) Adaptive EAP Estimation of Ability in a Microcomputer Environment. Appl Psychol Meas 6:431–444
31. Lugtmeijer S et al (2021) Are visual working memory and episodic memory distinct processes? Insight from stroke patients by lesion-symptom mapping. Brain Struct Funct 226:1713–1726

Figures

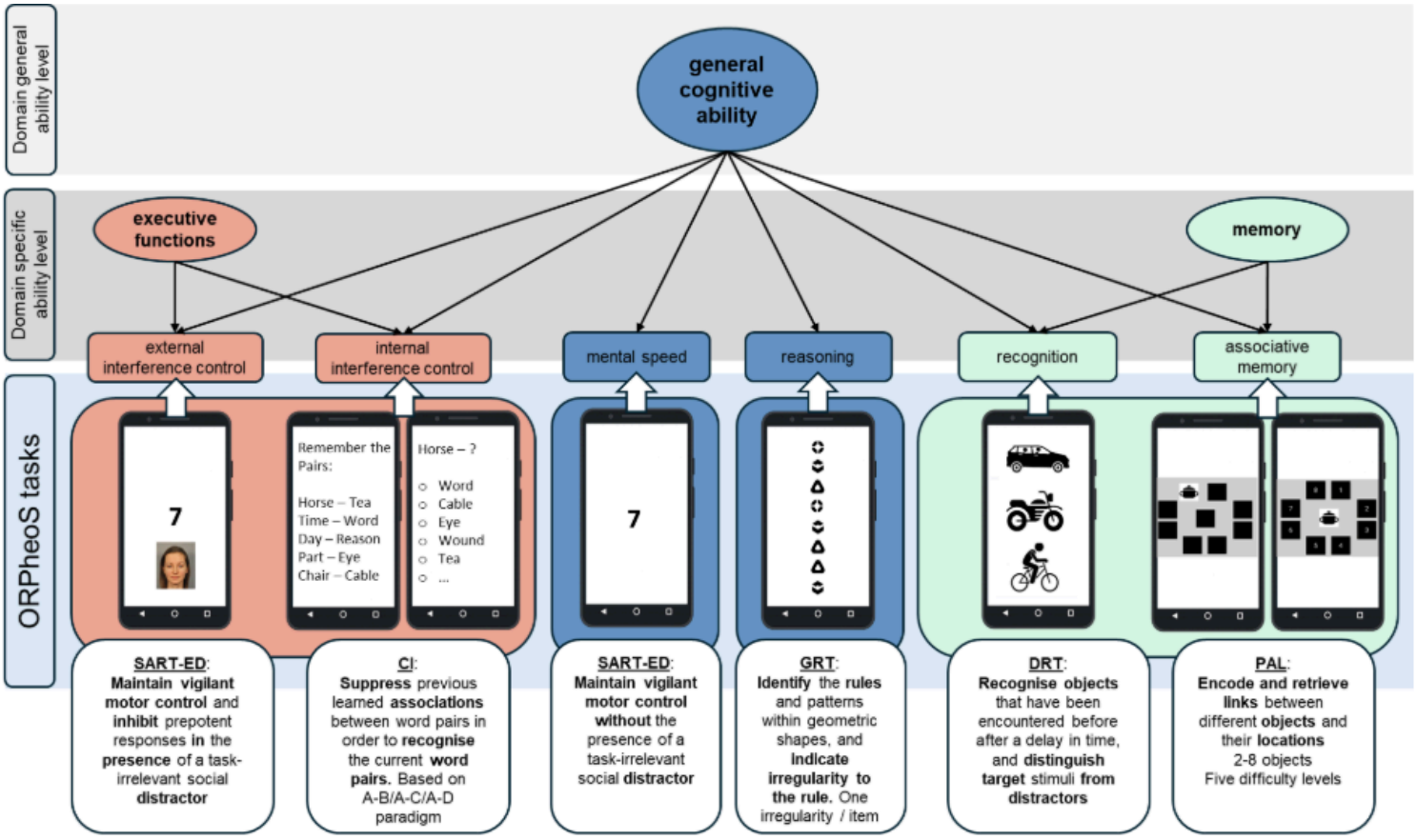


Figure 1

Theoretical taxonomic order of the ORPheoS task battery. Targeted domains are divided into executive function (left), memory (right), and general cognitive ability. Executive function measurement includes the SART-ED (Sustained Attention to Response Task with External Distraction; measures external interference control, based on the Go-No-Go paradigm) and the CI task (Cognitive Interference task; measures internal interference control; based on A-B/A-C/A-D paradigm). Reasoning (operationalized by the Geometric Reasoning Task, GRT) and mental speed (operationalised by the SART-ED "Go" trials without a distractor) serve as markers for general cognitive functioning. Memory comprises the DRT (Delayed Recognition Task of objects; measures recognition after incidental encoding) and the PAL (Paired Association Learning task; measures associative memory). The path diagram reflects a bifactor S-1 [28] type structural representation, illustrating both digital interfaces and the underlying taxonomy of the measured abilities.

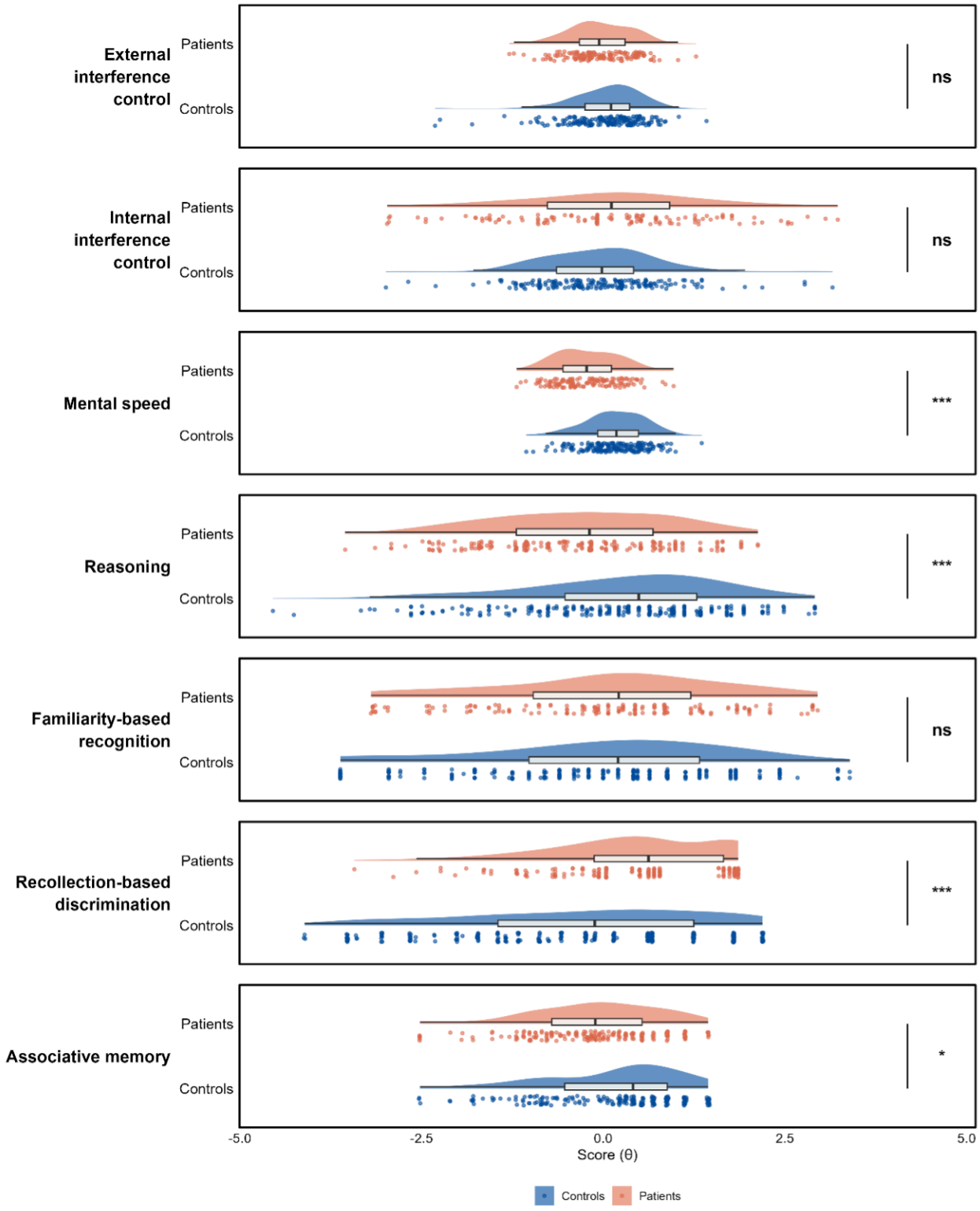


Figure 2

Latent ability (θ-score) estimates across cognitive domains by group. Raincloud plots display the empirical probability density (half-violins) and θ-scores for healthy controls (blue rain) and stroke participants (red rain). Boxplots denote the median and interquartile ranges. Statistical significance derived from independent samples *t*-tests are indicated adjacent to each plot (*ns = not significant, * p < 0.05, ** p < 0.01, *** p < 0.001).*



Figure 3

Correlation matrix and reliability estimates across cognitive domains measured by ORPheoS. This figure illustrates the relationships among the θ-scores and MoCA in the stroke sample. The upper right triangle displays the Pearson correlation coefficients (*r*) for the total sample, alongside sub-group correlations for controls and stroke participants. The lower left triangle displays corresponding scatterplots with overlaid linear regression fits. The diagonal displays the split-half reliability estimates for each task, with the exception of the MoCA, which is represented by a density plot of the observed scores. All *p*-values associated with the correlation coefficients were adjusted for multiple comparisons (*NA = not applicable as MoCA was not administered in the control group, * p < 0.05, ** p < 0.01, *** p < 0.001).*

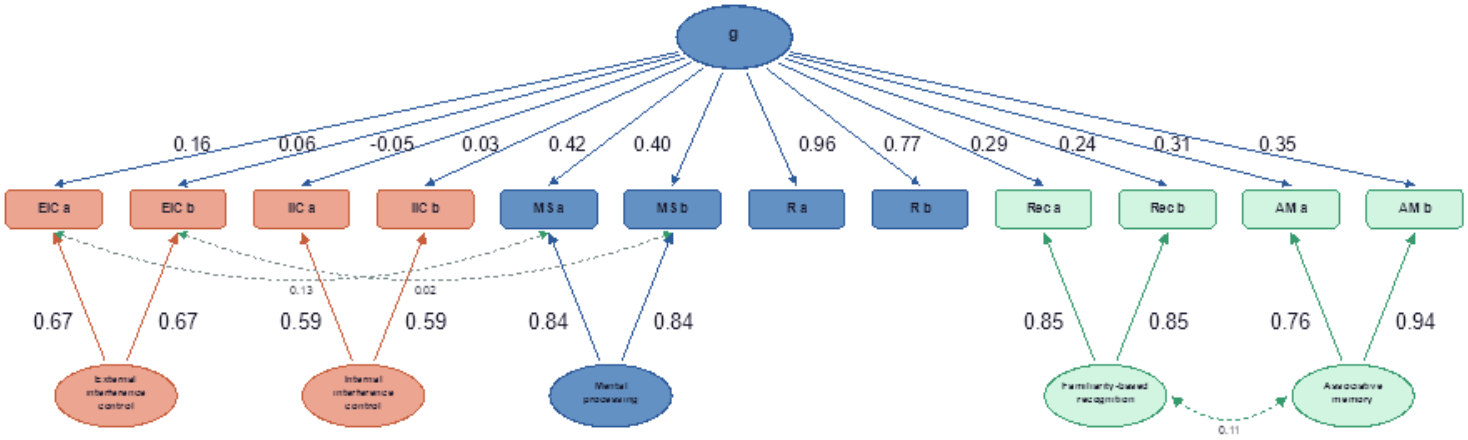


Figure 4

S-1 Bifactor model of cognitive abilities as measured by ORPheoS. The path diagram illustrates the latent structure of parcel-level θ-scores. The general cognitive factor (*g*, top) accounts for shared variance across all tasks, reflecting global cognitive ability, with geometric reasoning acting as the reference domain. Five specific factors (bottom) capture additional domain-specific variance, independent of *g*. Directed paths represent standardised factor loadings; rectangles represent θ-scores estimated from half-task item parcels (odd/even or interleaved). A dashed bidirectional arc indicates a covariance between the latent associative memory and familiarity-based recognition specific factors, accounting for local dependence per design. Additional residual covariances between corresponding theta scores of the EI and MS scores (parcel a with parcel a; parcel b with parcel b) were included to account for task-specific shared method variance. EIC: External Interference Control as derived from the SART-ED task; IIC: Internal Interference Control as derived from the CI task; MS: Mental Speed as derived from the SART-ED task; R: Reasoning as derived from the GRT; Rec: Recollection-based Recognition as derived from the DRT (without recollection-based discrimination items); AM: Associative Memory as derived from the PAL; a vs b is indicating the odd/even item parcels used for theta estimation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx