International Platform of Registered Systematic Review and Meta-analysis Protocols

INPLASY

Psychometric properties of digital cognitive assessments across the cognitive decline continuum: an umbrella review protocol

INPLASY202660108
doi: 10.37766/inplasy2026.6.0108
Received: 22 June 2026
Published: 22 June 2026

Corresponding author:
Joshua Panesar

j.r.panesar@student.rug.nl

Author Affiliation:
University of Groningen, Groningen, The Netherlands.

Panesar, JR; Wilschut, TJ; van Rijn, H

ADMINISTRATIVE INFORMATION

Support - NWO Take-off phase 2.

Review Stage at time of this submission - Preliminary searches.

Conflicts of interest - JP, TW, and HvR are based at the University of Groningen and are also associated with Precision Cognition Labs, a commercial partner involved in this project that develops SGMA, a digital memory test, for which the authors have an interest in validation. This interest is explicitly declared. As mitigation, the eligibility criteria, outcomes, and appraisal approach are pre-specified in this prospectively registered protocol, before screening begins; should the tool be covered by any review included in this umbrella review, it is screened, extracted, and appraised under the same pre-specified procedures as every other tool, independently and in duplicate, with no preferential weighting or interpretation; and all such decisions are documented so that they can be audited. All other authors declare no competing interests.

INPLASY registration number: INPLASY202660108

Amendments - This protocol was registered with the International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) on 22 June 2026 and was last updated on 27 July 2026.

INTRODUCTION

Review question / Objective This umbrella review aims to systematically map and synthesise evidence from existing review articles (systematic reviews, meta-analyses, scoping reviews, umbrella reviews, reviews-of-reviews, narrative reviews, and literature reviews) on the psychometric properties of digital cognitive assessment tools designed for unsupervised self-administration outside the clinical setting, across the cognitive decline continuum (see Item 10). A memory component is not an eligibility criterion and is applied at data extraction, where it determines which tools enter the tool-level maps (see Items 13 and 19).

The review will address the following questions:

Where does the existing review evidence sit methodologically across the continuum? That is, which continuum stages are well covered by methodologically high-quality reviews, and where are the methodological coverage gaps?

Which digital cognitive assessment tools designed for unsupervised self-administration have been evaluated in the review literature with reported psychometric properties, and at which stage(s) of the continuum?

What psychometric evidence (reliability, convergent validity, and diagnostic accuracy) has been reported for each tool, at which stage(s)?

Which tools have the strongest psychometric evidence reported in the review literature for a given stage of the continuum, and where are the most important gaps in that reported evidence?

The principal outputs are evidence maps that display, across the continuum, the methodological coverage of the review evidence and the reliability, convergent validity, and diagnostic accuracy reported for individual tools (Item 22).

Rationale Digital, home-administrable cognitive assessments have proliferated rapidly over the past decade as a solution to the scalability and access limitations of in-clinic neuropsychological testing, and as a possibility to detect subtle cognitive change earlier in the cognitive decline continuum than conventional instruments often allow. As the number of available tools has grown, so has the number of reviews evaluating them: Several systematic, scoping, and narrative reviews address overlapping sets of digital cognitive tests, populations, and measurement properties. These reviews differ in how they define "digital," which stages of the continuum they cover, which psychometric properties they report, and which reference standards they accept. In addition, they frequently include the same primary studies. The result is a fragmented evidence base of diagnostic test accuracy studies, which makes it difficult to determine, for any given tool, what is actually known about its reliability, validity, and diagnostic accuracy, and at which stage of the continuum that evidence applies.

To the best of our knowledge, no umbrella review has yet synthesised the systematic-review-level evidence on the psychometric properties of these tools across the continuum. Existing reviews are predominantly organised by tool or by technology type rather than by stage of the continuum, and none maps the psychometric evidence for individual tools against the stage of disease in which each tool has been validated. The existing reviews on digital cognitive assessment across the cognitive decline continuum mostly focus on a segment of the continuum, a single psychometric property, and/or review a specific subgroup of the digital cognitive assessments. Moreover, no review has established the methodological quality of the underlying review literature. Where multiple reviews report on the same tool, their estimates have not been reconciled, and the degree of primary-study overlap between them has not been quantified.

An umbrella review is the appropriate design for this problem because the unit requiring synthesis is the review rather than the primary study. By appraising and synthesising the existing reviews rather than re-extracting primary data, this review aims to make explicit where the review literature on these tools is methodologically sound and where it is weak, internally inconsistent, or absent. Moreover, it aims to establish, for each digital cognitive tool, what psychometric evidence has been reported in the review literature and at which stage of the continuum, including which tools have the strongest evidence reported for a given stage. This review, therefore, contributes both: a methodological map of where the field's synthesised evidence is sound, weak, contested, or missing, and a map of the psychometric evidence reported in the review literature for individual tools across the continuum. The two are complementary rather than independent outputs: what the reviews report at the tool level is the material from which the state of the synthesised evidence is characterised, and together they make explicit where new primary research or new reviews are most needed.

Condition being studied Cognitive function across the cognitive decline continuum, from unimpaired cognition through to dementia.

The operative criterion is progressive cognitive decline: populations are within scope where the cognitive impairment addressed is progressive in nature, regardless of the underlying condition. Cognitive impairment that is static or non-progressive in origin, for example traumatic brain injury, concussion, stroke without subsequent cognitive decline, developmental or perinatal conditions, ADHD, or autism, is outside the scope of this review.

The Alzheimer-spectrum stages are ordered by increasing cognitive impairment attributable to Alzheimer's disease, and are operationalised as categorical stages following the NIA-AA framework (1).

1. Healthy, cognitively unimpaired: cognitively unimpaired older adults, healthy aging, healthy controls, and preclinical AD (biomarker-positive but cognitively unimpaired, NIA-AA clinical stage 1)
2. Subjective Cognitive Decline: subjective complaints with normal objective performance, including subtle cognitive change not meeting MCI criteria
3. Mild Cognitive Impairment: objective impairment not meeting dementia criteria, including MCI due to AD and prodromal AD
4. Alzheimer's dementia: including early-stage AD

Cognitive decline arising from other progressive conditions is also within scope, including non-Alzheimer's dementia (vascular, Lewy body, frontotemporal, and mixed) and other progressive conditions in which cognitive decline is assessed, for example Parkinson's disease, Huntington's

disease, and multiple sclerosis. Because the clinical presentation at a given point on the continuum does not always determine the eventual aetiology, no grouping or display of these populations relative to the Alzheimer-spectrum stages is pre-specified. How they are grouped and presented in the synthesis is determined after data extraction, once the volume and distribution of the available evidence are known, and any such decision is reported transparently. Whether psychometric properties observed in non-Alzheimer's populations differ from those reported in AD is treated descriptively in the discussion. No a priori expectation of differential performance is built into the synthesis.

METHODS

Search strategy A five-block Boolean search strategy was developed combining (Block 1) digital technologies AND (Block 2) cognitive assessment tools AND (Block 3) psychometric properties AND (Block 4) cognitive impairment continuum populations AND (Block 5) review-type publication filter. Each block combines free-text title/abstract terms (with truncation and phrase searching) and database-specific controlled vocabulary. Filters applied across all databases: English-language, publication years 2006–2026.
Searches were run on the following dates with the following result counts:
– PubMed – 213 records (4 May 2026)
– Scopus – 156 records (5 May 2026)
– MEDLINE via EBSCOhost – 141 records (5 May 2026)
– APA PsycINFO via EBSCOhost – 89 records (5 May 2026)
– Web of Science Core Collection – 140 records (6 May 2026)
– Cochrane Library – 79 records (6 May 2026)
The complete Boolean search string for each database, with its interface, controlled-vocabulary mapping, applied filters, exact run time, and result count, has been recorded in full and will be reported with the published review in accordance with PRISMA-S.

Participant or population Adults aged 18 years or older across the cognitive decline continuum: cognitively unimpaired/healthy aging older adults, individuals with subjective cognitive decline (SCD), individuals with mild cognitive impairment (MCI), prodromal Alzheimer's disease, early-stage dementia, clinical Alzheimer's dementia, mixed dementia, vascular dementia, Lewy body dementia, or frontotemporal dementia.

The operative criterion is progressive cognitive decline. Reviews addressing cognitive decline in other progressive conditions, for example Parkinson's disease, Huntington's disease, and multiple sclerosis, are eligible where the review is clearly about cognitive decline and its assessment.

Reviews focused exclusively on populations in which the cognitive impairment is static or non-progressive in origin are not eligible, for example schizophrenia, traumatic brain injury, concussion, stroke without subsequent cognitive decline, developmental or perinatal conditions, ADHD, autism, or pediatric populations. Reviews of a condition otherwise within scope that do not address cognitive decline, for example a review of motor outcome measures in Parkinson's disease, are also not eligible.

Mixed reviews (those covering both eligible and ineligible populations, or both digital and non-digital tools) are retained at the title/abstract stage. Whether population-specific and digital-tool-specific psychometric data can be separately extracted is assessed at the full-text screening stage rather than at title/abstract screening.

Intervention This is a psychometric-properties umbrella review rather than an intervention umbrella review. The "index" of interest is digital cognitive assessment tools that satisfy all of the following:

(a) Unsupervised self-administration. Designed for, or validated in, fully unsupervised self-administration outside the clinical setting. Remote is not the same as unsupervised: tools requiring real-time human supervision are not eligible, whether that supervision is delivered in person or at a distance. This excludes videoconference-administered testing (tele-neuropsychology), telephone-administered cognitive batteries, and any test in which an examiner presents stimuli, times responses, or scores performance live. Tools that require in-person clinical administration, with no unsupervised home-based component described or validated, are not eligible.

(b) Consumer device. Delivered on a device the participant already owns (computer, tablet, or smartphone). Tools requiring hardware not ordinarily present in the home, such as virtual reality headsets, laboratory eye-trackers, or EEG, are not eligible.

(c) Active, performance-based cognitive test. Stimuli are presented to the participant and the participant's responses are scored. Passively

collected digital measures that require no cognitive task performance, such as keystroke dynamics, gait, sleep, phone-usage patterns, facial expression, driving telemetry, and comparable sensor-derived measures, are not eligible. Speech- and language-based measures are eligible where the participant performs a scored cognitive task, for example automated story recall, and are not eligible where speech is analysed only as a passively recorded signal.

"Digital" includes web-based and tablet-based batteries, smartphone applications, computerised tests, ecological momentary assessment, gamified or serious-game tasks, and any other electronically delivered cognitive assessment meeting (a) to (c). Paper-and-pencil tests and in-person clinic-administered tests with no digital component are excluded.

This is digital assessment rather than digital intervention. Reviews covering digital tools only as an intervention (cognitive training, rehabilitation, serious games for enhancement), where no digital tool is used to measure cognition, are not eligible.

A memory component (any task measuring encoding, retention, or retrieval of information, including delayed recall, recognition, associative/paired-associate learning, episodic memory, spatial memory, working memory, and prospective memory tasks) is not an eligibility criterion. It is applied at data extraction, where it governs which tools populate the tool-level maps rather than which reviews enter the review base (see Items 19 and 20).

Comparator Variable depending on the psychometric property under evaluation, and not a primary structuring axis of this review:

- For diagnostic accuracy outcomes (sensitivity, specificity, AUC): the reference standard used in the underlying primary studies, typically a clinical diagnosis of MCI or AD (NIA-AA, DSM-5, or equivalent criteria) and/or a comprehensive neuropsychological battery.
- For criterion or convergent validity: established cognitive assessments such as the MoCA, MMSE, or established neuropsychological batteries.
- For test–retest reliability: no external comparator (a within-instrument property; see Item 18).

Study designs to be included Review-type publications: systematic reviews, meta-analyses, scoping reviews, umbrella reviews, reviews-of-reviews, narrative reviews, and literature reviews. Excluded: primary empirical studies, case reports, case series, editorials, letters, commentaries, opinion and perspective pieces, conference abstracts, conference proceedings papers, consensus statements, clinical guidance documents, book chapters, and review protocols without results reported.

Eligibility criteria Eligibility is specified against the PICOS elements. Because this is an umbrella review of psychometric properties rather than an intervention review, the Comparator element is not used as an eligibility criterion. Each criterion is applied at one of two screening stages: title/abstract (T/A) or full text (FT). The stage is determined by whether the judgment can be made from the title and abstract alone or requires the full text (Item 20). Criteria requiring assessment of whether population-specific or digital-tool-specific psychometric data are separately extractable are applied at the full text, because that judgment requires the complete paper.

Title/abstract screening. A record is excluded at this stage only where one of the criteria below applies. Mixed reviews are retained, and their separability questions are deferred to the full text. Where a reviewer is unsure, the record is marked Maybe for discussion and is not excluded.

Population (Item 12). Include reviews covering at least one of (a) healthy or cognitively unimpaired adults studied in an ageing, dementia-risk, or cognitive-screening context, or (b) the cognitive decline continuum, even partially: healthy ageing / cognitively unimpaired older adults, SCD, MCI, prodromal/preclinical AD, AD dementia, or non-AD dementia (vascular, Lewy body, frontotemporal, mixed). Biomarker-positive preclinical AD (NIA-AA clinical stage 1) is cognitively unimpaired and belongs under (a). The operative criterion is progressive cognitive decline: a review is included where the cognitive impairment it addresses is progressive in nature, regardless of the underlying condition, so reviews addressing cognitive decline in other progressive neurological conditions, such as Parkinson's disease, Lewy body disease, Huntington's disease, and multiple sclerosis, are included, provided the review is clearly about cognitive decline and its assessment. Exclude reviews in which the cognitive impairment addressed is static or non-progressive in origin: traumatic brain injury, concussion, stroke without subsequent cognitive decline, developmental or perinatal conditions, ADHD, autism. Exclude reviews of a condition otherwise within scope that do not address cognitive decline, for example a review of motor outcome measures in Parkinson's disease. Retain mixed-population reviews; the

separability of population-specific data is assessed at full text.

Intervention / Index (Item 13). Include reviews covering at least one digital cognitive assessment tool that is designed for, or validated in, fully unsupervised self-administration outside the clinical setting, on a consumer device the participant already owns (computer, tablet, or smartphone). Remote is not the same as unsupervised: exclude reviews in which every reviewed tool requires real-time human supervision, whether that supervision is delivered in person or at a distance, including videoconference-administered testing (tele-neuropsychology), telephone-administered cognitive batteries, and any test in which an examiner presents stimuli, times responses, or scores performance live. Exclude reviews in which every reviewed tool requires hardware not ordinarily present in the home: virtual reality headsets, laboratory eye-trackers, EEG. Exclude reviews in which all reviewed tools require in-person clinical administration with no unsupervised home-based component described or validated. The index must be an active, performance-based cognitive test, in which stimuli are presented to the participant and the participant's responses are scored: exclude reviews in which every digital measure is passively collected and requires no cognitive task performance, such as keystroke dynamics, gait, sleep, phone-usage patterns, facial expression, driving telemetry, and comparable sensor-derived measures. Speech- and language-based measures are included where the participant performs a scored cognitive task, for example automated story recall, and excluded where speech is analysed only as a passively recorded signal. Digital assessment, not digital intervention: exclude reviews covering digital tools only as an intervention (cognitive training, rehabilitation, serious games for enhancement) where no digital tool is used to measure cognition. Retain mixed digital/non-digital reviews; separability of digital-tool-specific data is assessed at full text.

Study design (Item 15). Include review-type publications: systematic reviews, meta-analyses, scoping reviews, umbrella reviews, reviews-of-reviews, narrative reviews, and literature reviews. A formal search strategy is not required at this stage. Exclude primary empirical studies, case reports, case series, editorials, letters, commentaries, opinion and perspective pieces, conference abstracts, conference proceedings papers, consensus statements, clinical guidance documents, book chapters, and review protocols with no results reported. Publications combining a review with primary empirical data, for example a review followed by validation of the authors' own tool, are retained at this stage; whether the review component is separately extractable is assessed at full text.

Language. Exclude publications not in English.

Deferred to full text and never used to exclude at this stage: the extractability of digital-tool-specific psychometric data, separability of stage-specific data, and full-text retrievability.

Full-text screening. Applied to reports carried forward from title/abstract. The test at this stage is conjunctive and applies to a single tool. A qualifying tool is one reviewed tool that is at once digital and self-administered without real-time human supervision, delivered on a consumer device the participant already owns, and an active performance-based cognitive test. A review is included only where at least one qualifying tool is present and at least one psychometric property is reported for that same tool. Criteria met by different tools within the same review do not satisfy the test. Where a reviewer is unsure, the report is discussed and is not excluded unilaterally.

Population (Item 12). Assess, in mixed-population reviews retained at T/A, whether population-specific data can be separately extracted and analysed. Exclude where eligible and ineligible populations are pooled inseparably throughout and no result can be attributed to the eligible population. Reviews addressing progressive cognitive decline in non-Alzheimer's conditions remain eligible.

Intervention / Index (Item 13). Confirm that at least one reviewed tool meets the index criteria in full; a tool meeting some but not all of them does not qualify. Assess, in mixed reviews, whether data for qualifying tools can be separated from data for tools that do not qualify. The operative contrast is qualifying against non-qualifying, not digital against non-digital: a review reporting its digital results as one pooled block spanning unsupervised, videoconference-administered, and virtual-reality tools is excluded unless a qualifying tool can be identified individually. Where the review's description of a tool is insufficient to classify it, the tool's primary validation source may be consulted for classification only, never for extracting psychometric data; tools that remain undeterminable are treated as non-qualifying.

Outcome (Item 18). Include reviews reporting at least one psychometric property for at least one

qualifying tool: reliability, validity, sensitivity, specificity, diagnostic accuracy, ICC, AUC, or comparable measurement-quality indicators. A review that does not present tool-specific psychometric data as such is included where those data can be recovered from its tables, figures, appendices, or subgroup analyses. Exclude reviews where all psychometric results are pooled inseparably and no figure, subgroup analysis, or narrative estimate can be attributed to a qualifying tool. The form in which a psychometric property is reported does not affect eligibility: a review reporting psychometric evidence qualitatively, as ticks, ratings, a rating matrix, or narrative judgements such as "adequate test–retest reliability," satisfies this criterion. The requirement for extractable numerical coefficients is a synthesis filter applied at extraction rather than an eligibility criterion; such reviews contribute to the review-level map, the supplementary tables, and the narrative synthesis, but not to the tool-level maps. This is distinct from the absence of any psychometric evaluation: reviews reporting only usability, feasibility, acceptability, uptake, adherence, or descriptions of the tools and how they are administered report no psychometric property and are excluded.

Study design (Item 15). Confirm the design against the full text. A report that proves on full reading to be a primary study, commentary, consensus statement, or other excluded type is excluded here, whatever its abstract indicated. Exclude review protocols with no results reported. Publications combining a review with the authors' own primary data are included where the review component has its own results and can be read independently of the primary study, and are excluded where the review serves only as an introduction to those data.

Full-text retrieval. Retrievability is a retrieval outcome rather than an eligibility exclusion. Full texts are sought via the RUG proxy, interlibrary loan, and direct author contact. A report not obtained is recorded in the PRISMA 2020 flow diagram under reports not retrieved, listed with its citation in the supplementary material, and does not enter the assessed-for-eligibility count.

Comparator (Item 14). Not a screening criterion at either stage. The comparator varies for each psychometric property (reference standard for diagnostic accuracy, reference instrument for convergent validity, none for reliability) and is recorded during data extraction rather than used to include or exclude reviews. Reported here for completeness only.

Additional criteria not defined by PICOS (justified per INPLASY Item 16 guidance). Language: English-language publications only. This restriction is applied for feasibility (the review team's working language) and is acknowledged as a potential source of language bias in the limitations. Time frame: publication years 2006–2026. The lower bound reflects the emergence of the modern computerised and home-based cognitive-assessment literature, as reviews of tools predating this period are unlikely to bear on currently deployable digital tools. The window is applied at the search stage and is not used as a post hoc exclusion. Country/setting: no restriction.

Specification of the criteria. The eligibility criteria were refined following pilot title/abstract screening (n = 50), agreed by all three reviewers, and applied prospectively from that point onward. The full-text criteria were likewise specified and agreed by the same reviewers before full-text screening began.

Information sources Six bibliographic databases:
– PubMed
– Scopus
– MEDLINE via EBSCOhost
– APA PsycINFO via EBSCOhost
– Cochrane Library
– Web of Science Core Collection
Database access is via the University of Groningen RUG proxy.
Search reporting will follow PRISMA-S.

Main outcome(s) This review has outcomes at two levels. At the review level, the outcome is the methodological quality and coverage of the eligible review literature across the continuum, appraised with ROBIS (Item 21) and reported as the review-level evidence map (Item 22). At the tool level, the outcomes are the psychometric properties reported for the included digital cognitive assessment tools, which populate the tool-level evidence maps (Item 22):
Reliability: primarily test–retest reliability, recorded as the intraclass correlation coefficient (ICC) or test–retest correlation (r) as reported in the included reviews. Where an ICC is reported, its form (model and type) is recorded where stated and flagged 'unspecified' otherwise (2).
Diagnostic accuracy: sensitivity and specificity (as a pair) against a clinical reference standard, and/or area under the receiver-operating-characteristic curve (AUC). Reference standards include a clinical diagnosis of MCI or AD (NIA-AA, DSM-5, Petersen, or equivalent criteria), a full neuropsychological battery, or a biomarker-supported diagnosis.
Convergent validity: Pearson or Spearman correlation (r) between the digital tool's score and

an established reference instrument, such as a global cognitive screener (the MoCA or MMSE) or an established neuropsychological battery.
Benchmarks for what constitutes adequate reliability, validity, and diagnostic accuracy will be informed by established psychometric and diagnostic guidance (e.g., COSMIN/Terwee (3,4); consensus diagnostic-accuracy criteria (5); AUC interpretation (6)) and are finalised at the analysis stage. Their application to the maps is described in Item 22.

Additional outcome(s) Beyond the psychometric outcomes specified in Item 18, the following descriptive characteristics will be extracted at two levels (further detailed in Item 20) and will feed the evidence maps and accompanying narrative synthesis (Item 22):
At the review level: citation, review type, databases searched, number of included primary studies, continuum stages covered, tools covered, psychometric properties reported, reference standards used, and ROBIS risk-of-bias rating. All included reviews contribute review-level data.
At the tool level: tool name (with cross-review tool-identity harmonisation, see Item 20), continuum stage of the population in which the tool was evaluated, sample size, cognitive domain(s) assessed (with particular attention to which memory subprocess is targeted, such as delayed recall, recognition, associative learning, spatial, working, or prospective memory), device platform (web/computer, tablet, smartphone), and administration time.
Tool-level extraction is restricted to qualifying tools that include at least one memory component, as defined in Item 13.
The review-level data populate the review-level methodological-coverage map. The tool-level psychometric data populate the tool-level psychometric-property maps and the descriptive supplementary tables (Item 22).

Data management Reference management and deduplication. All records from the six databases are imported into Rayyan. Duplicate candidate pairs are resolved manually before screening, and only confirmed duplicates are removed. The number of confirmed duplicates and the unique-record count entering screening are reported in the PRISMA 2020 flow diagram.

Pilot screening. Fifty abstracts were independently screened by all three reviewers before full title/abstract screening, to assess inter-rater agreement and refine the operationalisation of the criteria. The refined criteria were agreed by all three reviewers and applied prospectively from that point (see Item 16). Inter-rater agreement across the three reviewers is reported for the title/abstract stage.

Title/abstract screening. Three independent reviewers (the lead author and two co-screeners) screen all records in Rayyan, blinded to each other's decisions. Conflicts are resolved through discussion and persistent disagreements through adjudication.

Full-text screening. Dual screening by the lead author and one co-screener. Exclusions are documented with reasons (per the Stage-2 eligibility criteria in Item 16) for the PRISMA 2020 flow diagram. Where more than one exclusion criterion applies to a report, the clearest applicable reason is recorded as the primary reason for exclusion, agreed in discussion between reviewers.

Data extraction. A two-level extraction template is piloted on 2–3 anchor reviews and refined before full extraction. At the review level (one row per review): citation, review type, databases searched, number of primary studies, population and continuum stages covered, tools covered, psychometric outcomes reported, reference standards used, and ROBIS rating. At the tool level (one row per qualifying memory-component tool per stage per review): the harmonised tool name, the continuum stage of evaluation, sample size, the psychometric coefficients reported (test–retest ICC or r, sensitivity, specificity, AUC, convergent validity r, and the reference instrument), and whether qualifying-tool data are separately extractable in mixed reviews.

Stage attribution. Each extracted coefficient is attributed to a continuum stage using the stage of the sample as reported in the review or, where unstated, the stage fixed by the review's inclusion criteria. Coefficients derived from a contrast between two populations, such as sensitivity, specificity, and AUC for discriminating one group from another, are attributed to the target condition, with the comparison group recorded alongside. Coefficients that cannot be attributed to a single stage, including those pooled across populations, are recorded as such together with the populations they span; how these are presented in the synthesis is determined at the analysis stage (Item 22).

Tool-identity harmonisation. The same tool is often referred to by different names across reviews. A tool-identity list is built during pilot extraction and updated throughout the process, with the lead author and one co-reviewer independently classifying each new name and resolving

disagreements through discussion. Decisions about which reviews contribute to which maps, and how discrepant estimates are reconciled, are described in Item 22.

Software. Rayyan (screening and deduplication), a spreadsheet-based extraction template, ROBIS, CCA calculation in a spreadsheet or R, and figure preparation in R or equivalent.

Quality assessment / Risk of bias analysis The risk of bias of each included review is appraised using ROBIS (7) as the only risk-of-bias tool. ROBIS is applied in its three phases: assessing relevance to the review question; identifying concerns across four domains (study eligibility criteria; identification and selection of studies; data collection and study appraisal; and synthesis and findings); and reaching an overall judgement of low, unclear, or high risk of bias. Each included review is appraised independently and in duplicate by two reviewers, with disagreements resolved by discussion and, where needed, adjudication by a third reviewer.
ROBIS is selected because it is a validated review-level instrument designed for systematic reviews of any type, including reviews of diagnostic accuracy, and explicitly names overviews of reviews among its target uses. Its low/unclear/high judgement maps directly onto the methodological dimension of this review's outputs, so a single appraisal output informs both the appraisal and the synthesis.
Risk of bias does not affect inclusion: no review is excluded on the basis of its ROBIS rating. Quality is instead incorporated into the synthesis, as a dimension of the review-level map, as a risk-of-bias encoding on the tool-level maps, and as a sensitivity analysis restricting the synthesis to low-risk-of-bias reviews (Item 24). Overlap between reviews at the primary-study level is quantified separately using the Corrected Covered Area method (8).
A single review-level tool is used deliberately, as there is currently insufficient empirical evidence to prefer one review-appraisal tool over another (9). AMSTAR 2 and QUADAS-2 are not used as the primary tools: AMSTAR 2 is intended for reviews of healthcare interventions (10), and QUADAS-2 appraises primary diagnostic-accuracy studies rather than reviews. Where included reviews report their own QUADAS-2 or GRADE assessments, these are extracted and presented in the supplementary tables as context. Formal GRADE/GRADE-DTA is not applied, as the umbrella-review design and the heterogeneity of tests, populations, reference standards, and thresholds across the included reviews preclude a single meaningful certainty rating. Methodological confidence is instead conveyed through the ROBIS risk-of-bias ratings, reflected in the synthesis and in the low-risk-of-bias sensitivity analysis (Item 24), with each review's own certainty statements reported narratively.
Reporting of this umbrella review follows PRIOR (11).

Strategy of data synthesis A narrative synthesis is planned, structured around the cognitive decline continuum and organised into two complementary units of analysis: the review (to characterise the methodological evidence landscape) and the tool (to summarise the psychometric evidence reported in the review literature for each included assessment). The synthesis is anchored by evidence maps spanning the continuum stages defined in Item 10.

We intend to visualise the psychometric evidence reported for the included tools across the continuum in separate maps for reliability, convergent validity, and diagnostic accuracy (sensitivity/specificity), complemented by a review-level map summarising the methodological coverage and risk of bias of the underlying review evidence. One plausible design plots a psychometric coefficient (for example, the ICC) on the vertical axis and the continuum on the horizontal axis, with tools shown as individual points and the strength of the supporting evidence encoded visually (for example, through point opacity). The final form of each map, including the axes, the unit summarised per cell, the grouping and display of populations, and any benchmarks used to aid interpretation, will be determined after data extraction, once the number of contributing tools and the sparsity of the evidence are known, and any such decisions will be reported transparently.

For each tool, the reported psychometric properties are summarised qualitatively per continuum stage, drawing on both the coefficient-based evidence and the qualitative or narrative judgements from reviews that do not enter the tool-level maps. Discrepant estimates across reviews are reported together with their range and the ROBIS ratings of the source reviews. The narrative synthesis is structured stage by stage, each stage concluding with a summary of which tools have the strongest psychometric evidence reported in the review literature and where the gaps lie.

Given the heterogeneity of reporting formats and reference standards across the included reviews,

the tool-level maps are anticipated to be sparsely populated in places, particularly for diagnostic accuracy and convergent validity and at the less well-studied points of the continuum. Empty cells are interpreted as gaps in the review evidence rather than as evidence of poor tool quality, and this is made explicit in the figure captions and discussion. The feasibility of the proposed maps is confirmed during a pre-extraction pilot on 2–3 anchor reviews; where a property is too sparsely populated to support a useful figure, it may be presented as a supplementary table, reported transparently.

No quantitative meta-analysis is planned. The heterogeneity of tools, stages, reference standards, and reporting formats across reviews, combined with the umbrella-review design, makes statistical pooling neither feasible nor appropriate. Heterogeneity is therefore described narratively rather than quantified (no I^2 or τ^2), and overlap among reviews at the primary-study level is quantified via the Corrected Covered Area method (Item 21). Where the same primary studies underlie estimates reported by more than one review, this is taken into account when summarising the evidence for a tool, so that repeated reporting of the same underlying evidence is not presented as independent accumulation.

Subgroup analysis The primary stage- and tool-level stratification (the continuum stages and the per-tool structure of the synthesis) is the pre-specified structuring design of the synthesis, not a subgroup analysis. Beyond it, no confirmatory subgroup analyses are pre-specified. The following is exploratory and hypothesis-generating, undertaken only where the number of supporting reviews permits and reported transparently as post hoc: device platform (web/computer, tablet, smartphone). Review type and ROBIS rating are not treated as subgroups here, as both are addressed as sensitivity analyses (Item 24) and, for ROBIS, as a synthesis dimension.

Sensitivity analysis Sensitivity analyses test the robustness of the synthesis to the methodological quality and the type of the included reviews. As no quantitative pooling is undertaken (Item 22), they are applied to the descriptive and narrative synthesis rather than to a summary estimate: the synthesis is re-generated under each restriction and compared with the picture obtained from the full set of included reviews.
1. Restriction to reviews rated at low risk of bias on ROBIS. Where too few reviews are rated low risk for a meaningful comparison, the restriction is instead applied by excluding reviews rated at high risk of bias, and this is reported.
2. Restriction to systematic reviews and meta-analyses, excluding scoping, narrative, and literature reviews.
Each restriction is applied independently rather than cumulatively, and is undertaken where the number of contributing reviews permits. Where the picture changes meaningfully under either restriction, this is reported and discussed in the results and discussion sections. Any further sensitivity analysis undertaken after data extraction is reported transparently as post hoc.

Language restriction English only. Reviews not published in English will be excluded.

Country(ies) involved Netherlands.

Keywords Umbrella review; digital cognitive assessment; unsupervised self-administration; psychometric properties; cognitive decline continuum; memory; mild cognitive impairment; remote neuropsychological assessment; diagnostic accuracy.

Dissemination plans The results will be submitted for publication in a peer-reviewed journal in the field of cognitive aging or neuropsychology (candidate journals include Neuropsychology Review, Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, and Ageing Research Reviews), and presented at one or more national or international conferences in cognitive aging, neuropsychology, or digital health. Available as supplementary materials to the published manuscript: the PRISMA 2020 flow diagram, ROBIS ratings for each included review, the full search strings, the two-level extraction dataset (review-level and tool-level), the tool-identity harmonisation table, the count of reviews entering each map, the rationale for any review contributing to the review-level map but not to the tool-level maps, and the evidence maps (the review-level methodological-coverage map and the tool-level reliability, diagnostic accuracy, and convergent validity maps) at full resolution.

Contributions of each author

Author 1 - Joshua Panesar - CRediT: Conceptualisation; Methodology; Formal analysis; Investigation; Data curation; Writing – original draft; Writing – review & editing; Visualisation; Project administration; Software; Validation.
Email: j.r.panesar@student.rug.nl
Author 2 - Thomas Wilschut - CRediT: Conceptualisation; Methodology; Investigation; Writing – review & editing; Supervision; Validation.

Email: t.j.wilschut@rug.nl
Author 3 - Hedderik van Rijn - CRediT: Conceptualisation; Methodology; Writing – review & editing; Supervision; Resources; Funding acquisition.
Email: d.h.van.rijn@rug.nl

References

1. Jack CR, Andrews JS, Beach TG, Buracchio T, Dunn B, Graf A, et al. Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's Association Workgroup. Alzheimers Dement. 2024 Aug;20(8):5143–69. doi:10.1002/alz.13859
2. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016 Jun;15(2):155–63. doi:10.1016/j.jcm.2016.02.012
3. Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, De Vet HCW, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018 May;27(5):1147–57. doi:10.1007/s11136-018-1798-3
4. Terwee CB, Bot SDM, De Boer MR, Van Der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007 Jan;60(1):34–42. doi:10.1016/j.jclinepi.2006.03.012
5. Consensus Report of the Working Group on: "Molecular and Biochemical Markers of Alzheimer's Disease." Neurobiol Aging. 1998 Mar;19(2):109–16. doi:10.1016/S0197-4580(98)00022-0
6. Mandrekar JN. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. J Thorac Oncol. 2010 Sep;5(9):1315–6. doi:10.1097/JTO.0b013e3181ec173d
7. Whiting P, Savović J, Higgins JPT, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol. 2016 Jan;69:225–34. doi:10.1016/j.jclinepi.2015.06.005
8. Pieper D, Antoine SL, Mathes T, Neugebauer EAM, Eikermann M. Systematic review finds overlapping reviews were not mentioned in every other overview. J Clin Epidemiol. 2014 Apr;67(4):368–75. doi:10.1016/j.jclinepi.2013.11.007
9. Higgins JPT, Cochrane Collaboration, editors. Cochrane handbook for systematic reviews of interventions. Second edition. Hoboken, NJ: Wiley-Blackwell; 2019. 694 p. (Cochrane book series).
10. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017 Sep 21;j4008. doi:10.1136/bmj.j4008
11. Gates M, Gates A, Pieper D, Fernandes RM, Tricco AC, Moher D, et al. Reporting guideline for overviews of reviews of healthcare interventions: development of the PRIOR statement. BMJ. 2022 Aug 9;378:e070849. doi:10.1136/bmj-2022-070849