Check for updates

Rethinking neuropsychological test validity in dementia assessment: a critical review in the age of neuroimaging and digital markers

Seyul Kwak*

Department of Psychology, College of Social Sciences, Pusan National University, Pusan, Republic of Korea

Neuropsychological tests are essential tools for evaluating dementia and related neurocognitive disorders, with their clinical utility determined mainly by their validity. This paper critically reviews the diverse evidence supporting the validity of neuropsychological tests in dementia assessment. Criterion validity is discussed in relation to the tests' ability to predict clinical diagnoses and underlying brain pathology, with a focus on their sensitivity to functional impairments and progressive neuropathological changes. Construct validity is explored through the lens of cognitive processes underlying test performance, using evidence from correlation structures and experimental paradigms. Furthermore, the paper examines the impact of emerging digital technologies on the evaluation of neuropsychological test validity, highlighting contrasts with traditional validation methods. The review identifies discrepancies between different types of validity evidence, emphasizing the need to contextualize validity within specific clinical and research applications. By addressing the conceptual limitations and trade-offs between validation approaches, this study proposes a comprehensive framework for interpreting validity evidence. Ultimately, it offers theoretical and practical implications for enhancing the robustness of neuropsychological tests in clinical practice and research.

Introduction

Neuropsychological assessment includes tools for evaluating individuals with suspected dementia. These assessments provide evidence to determine the presence of cognitive impairments, differentiate observed cognitive decline from normal aging, assess the severity of dementia, and identify the underlying pathology causing the deficits. Neuropsychological tests aim to measure cognitive abilities and detect deficits associated with specific brain structures and functions. Over the course of their long history, neuropsychologists have systematically gathered psychometric evidence to validate neuropsychological tests and ensure they serve their intended purpose.

Validity, commonly defined as the degree to which a test measures what it claims to measure, is a central consideration in the development and application of psychometric instruments (American Educational Research Association et al., 2014; Cronbach and Meehl, 1955). The type of validity sought for a test depends on its intended purpose and context, and the methods for establishing validity are similarly influenced by these objectives (Chen, 2018).

Evidence supporting test validity is often categorized into two primary dimensions: (1) whether the test is practical and useful, and (2) what specific attributes the test possesses (Hughes, 2017). These dimensions correspond to criterion-based and construct-based validation approaches, representing two major frameworks of scientific evidence in test validation. Neuropsychological tests, like psychometric approaches in general, derive their validity from their clinical and research applications, with these two types of evidence forming the basis for validation.

While there is a broad consensus regarding the conceptual framework for validity, further discussion is required to address specific issues arising from contemporary science. Moreover, advancements in technological methodologies are challenging traditional clinical neuropsychological approaches, necessitating new validation methods. This review critically examines the various types of evidence supporting the validity of neuropsychological tests in the context of dementia and related neurocognitive disorders. It highlights the multiplicity of clinical criteria and construct-based evidence expected in these contexts and explores how traditional perspectives— such as empiricism and cognitivism—have been integrated into dementia assessment. Additionally, the review considers recent advancements in biomarker, neuroimaging, and digital marker research, examining how these developments have altered the landscape of validity evidence. Finally, it discusses potential discrepancies between different types of validity evidence, offering insights into their implications for clinical and research settings.

Criterion validity of clinical outcome

Criterion validity involves evaluating whether a psychological test captures practical and meaningful real-world criteria, such as daily functioning or occupational adjustment, to validate its use as a tool for clinical diagnosis (Kane, 2001). This approach represents a foundational principle in clinical neuropsychology and has historically been one of the most widely utilized methods for test validation, often considered the "gold standard" in psychometric evaluation (Anastasi, 1950). Furthermore, empirical studies on the validity of neuropsychological test batteries frequently focus on this type of validation (Pawlowski et al., 2013). In the following sections, we explore criterion validity by examining its temporal dimensions (i.e., concurrent or predictive) and levels of analysis (i.e., clinical outcomes or brain pathology).

Concurrent criterion validity

Concurrent criterion validity is the most rudimentary approach to criterion validation, where a psychological test is assessed against a criterion measured at the same point in time (American Educational Research Association et al., 2014). This type of validity relies on practical judgment criteria grounded in the specific and practical purposes of the test, allowing its utility to be verified without heavy reliance on theoretical constructs.

For many individuals undergoing neuropsychological assessments, two primary objectives emerge: (1) determining whether cognitive impairments are present due to a neurological condition, and (2) establishing whether these impairments currently affect the

individual's ability to perform daily activities. Clinical diagnoses summarize comprehensive judgments about changes in daily functioning caused by cognitive impairments, serving as a robust reference target for assessing the utility of testing tools. For example, the diagnostic criteria for dementia emphasize difficulties in independent daily functioning (McKhann et al., 2011), and neuropsychological measures of cognitive function are validated as reliably predicting these outcomes (Farias et al., 2003; Fields et al., 2010; Jefferson et al., 2006).

It should be noted, however, that two distinct aspects— neurological condition and functioning level—are intermingled within the criterion of clinical diagnosis, which in turn serve as separate benchmarks for validating neuropsychological tests. First, a primary criterion for test validity is the test's ability to detect the organicity of impairment that indicates a neurological condition. In other words, the tests are deemed sensitive if the measured property responds to the presence of a biological cause of a clinical syndrome. Until the widespread availability of neuroimaging techniques in the 1980s, neuropsychological tests were predominantly used to identify the presence and track the location of brain injuries. At the time, validation relied on autopsy findings or neuroimaging as reference criteria, which formed the basis of tests as practical proxies for brain function measurement when direct assessment of neural damage was unavailable. For instance, a cluster of cognitive deficits frequently observed in patients with frontal lobe damage was referred to as "frontal lobe syndrome," leading to the development of tests specifically designed to assess "frontal lobe functions" (Dubois et al., 2000; Malloy and Richardson, 1994). These traditional frameworks of the neuropsychological battery have been utilized to differentiate neurological conditions, which remain foundational to the development and composition of tests (Reitan and Wolfson, 2009).

Second, the functioning level implied in the clinical diagnosis also serves as another concurrent criterion for validity. For older adults with dementia, maintaining independent daily activities is directly linked to their quality of life and the burden experienced by caregivers. In cases where caregivers are unavailable to provide assessments of daily functioning, neuropsychological tests become a critical source of clinical inference. Tests measuring episodic memory and executive function consistently exhibit strong correlations with levels of daily functional capacity (Fields et al., 2010; Overdorp et al., 2016). Furthermore, neuropsychological tests often predict the degree of functional impairment more accurately than biomarkers associated with neurological conditions (Kwak et al., 2021b).

While the criterion of brain damage or functional impairment provides a foundational framework for evaluating test utility, this approach has inherent limitations, as the meaning of the criterion itself can be enumerated. For example, using "frontal lobe damage" as a criterion for test validation is problematic as an underspecified definition because individual patients exhibit substantial heterogeneity in the location, pattern, and extent of brain injuries. Consequently, tasks designed to measure specific brain functions (e.g., frontal lobe tests) are prone to misidentifying deficits in patients with frontal lobe damage (Demakis, 2004; Phillips, 2004).

Similarly, using clinical diagnoses as a criterion can raise other issues. While diagnostic categories often provide a convenient and practical reference for validating test utility, they may prioritize symptom description over the underlying disease entity or etiology. Because the characteristics of a disorder's symptoms often result from

complex and cumulative interactions of risk factors, the underlying disease stages, shared neuropathologies, and deficits in specific cognitive processes may remain unspecified. Even when a test demonstrates excellent diagnostic classification, its relationship to the biological or psychological construct it purports to measure may be unclear, leaving its interpretability vulnerable to alternative, unanticipated factors (Strauss and Smith, 2009). Consequently, while the evidence of diagnostic validity is suggestive of a test's practical utility, the intrinsic validity of the criterion itself will remain largely unevaluated. For the advancement of operationally defining the diagnosis of dementia, researchers have advocated for separating clinical "disorders" from biologically defined "diseases" in diagnostic frameworks (Jack et al., 2024). However, this distinction has not yet been applied in the common practice of test validation.

Another limitation of concurrent validity is its reliance on criterion characteristics at a single time point, failing to account for changes over time. For instance, the DSM-5 diagnostic criteria for mild and major neurocognitive disorders emphasize objective cognitive decline as a core requirement. This reliance on test results during clinical diagnosis introduces the risk of criterion contamination, wherein the validity criterion itself is influenced by the test outcomes (Noel-Storr et al., 2014; Weuve et al., 2015). Such contamination risks exaggerate a test's diagnostic utility through circular reasoning. When test validation studies lack pre-planned designs, information gathered from patients may inadvertently influence clinical criteria through indirect pathways. Even if the diagnosis is not explicitly contaminated by the leakage of target test results, as with mild cognitive impairment (MCI), diagnosis criteria derived from similar cognitive tests can still introduce inherently shared elements between the test under validation and the reference standard.

Predictive criterion validity

One approach to maximizing the practical utility of criterion validity and addressing the limitations of concurrent validity is to evaluate predictive validity. Predictive validity assesses whether a psychological test provides information that can forecast future outcomes, distinguishing it from concurrent validity, which evaluates alignment with present criteria (American Educational Research Association et al., 2014). By using clinically meaningful benchmarks, predictive validity shares similarities with concurrent validity but emphasizes the test's ability to predict future clinical outcomes. This approach mitigates issues of criterion contamination present in concurrent validation by prospectively evaluating whether a test can predict future clinical impairments.

Accumulating research highlights the significant role of neuropsychological tests in providing critical insights into the progression of dementia symptoms. Scores obtained through neuropsychological tests can predict the likelihood of dementia onset or rapid cognitive decline years in advance (Bäckman et al., 2005; Belleville et al., 2014; Chapman et al., 2011; Jang et al., 2017). While comprehensive neuropsychological tests are more time-intensive than screening tools, such as the Mini-Mental State Examination (MMSE), they offer superior accuracy in distinguishing mild cognitive impairment (MCI) patients who are at risk of developing dementia (Kim et al., 2017). Numerous studies further indicate that

neuropsychological assessments enhance diagnostic accuracy and provide additional information about longitudinal changes in daily functioning, particularly for conditions of MCI and stroke (Donders, 2020). Thus, even within the same diagnostic category, neuropsychological tests contribute valuable explanatory power regarding future clinical outcomes.

It is also notable that repeated testing in follow-up assessments improves predictive validity by capturing changes over time. For example, one study has shown that older adults diagnosed with MCI and subsequently re-evaluated using MMSE or clinical rating scales showed dubiously low reversion rate from MCI to normal in the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (Thomas et al., 2019). In other words, a higher rate of MCI-to-Normal conversion implies the initial diagnosis of MCI was unreliable and less meaningful, which was the case when using a less valid cognitive test set. However, this artificial diagnostic bias could be adjusted by follow-up assessments incorporating neuropsychological tests, particularly those measuring episodic memory (Thomas et al., 2019). In another similar study, subtle cognitive changes observed in neuropsychological tests for 12 months have been shown to predict dementia onset with high accuracy over subsequent follow-ups (18–120 months) (Nation et al., 2019). These predictions often surpass the prognostic capabilities of major dementia biomarkers, highlighting the added value of neuropsychological tests. This series of studies demonstrates that repeated measurements of neuropsychological assessments enhance sensitivity to neurological changes and disease progression.

Both predictive and concurrent validity rely on clinically meaningful benchmarks, but the time point of assessment influences the validity evidence obtained. For instance, a study compared how a brief screening test and comprehensive tests correlate with current and future brain structural atrophy (Kim et al., 2017). The result showed that in patients already diagnosed with dementia or major neurocognitive disorders, MMSE captures the progression of neuropathology as effectively as comprehensive neuropsychological tests (Kim et al., 2017). Conversely, comprehensive neuropsychological batteries were uniquely accurate in detecting subtle future neuropathological changes in individuals with mild cognitive impairments. Furthermore, while hallmark clinical symptoms of Alzheimer's disease (AD) often involve deficits in long-term memory, preclinical stages (a decade or more before onset) are frequently characterized by earlier changes in immediate memory (i.e., learning), processing speed, and executive functions (Amieva et al., 2014; Bilgel et al., 2014; Mortamais et al., 2017; Younes et al., 2019). In evaluating dementia risk, it is crucial to distinguish between concurrent clinical symptoms and preclinical signs, as these may represent different phenomena.

Another consideration when reviewing predictive and concurrent validity is that, in real clinical settings, objective tests interact in complex ways with subjective reports (patients' subjective cognitive complaints, metamemory, and caregivers' informant-rated observations of functioning), such that the apparent characteristics of validity may be modulated. Primarily, subjective measures of cognition contain predictive signals that are not captured by neuropsychological tests (Jessen et al., 2014; Lai, 2014; Perrotin et al., 2015). In addition, subjective reports can index contextual factors— such as anxiety/depression, insight, and environmental demands— that help recalibrate the clinical meaning of identical performance

scores. This moderation can be observed explicitly within statistical models: considering neuropsychological tests in isolation may fail to capture clinical validity in the strict sense, whereas combining patient and informant reports with test scores can enhance the utility of certain measures (Kwak et al., 2023). Moreover, other studies show that baseline subjective cognitive complaints, even when only weakly related to concurrent objective cognitive performance, are associated with distinct long-term trajectories over extended follow-up (Morrison and Oliver, 2023). Accordingly, incorporating subjective reports alongside neuropsychological measures in predictive models leaves room to improve the predictive validity of existing test scores.

Criterion validity of brain pathology

Distinguishing whether neural damage is present and whether cognitive problems reported by examinees are due to neurological conditions rather than psychiatric states has long been a fundamental goal of neuropsychological testing (Bilder, 2011). Many neuropsychological tests have been validated based on their sensitivity to detecting brain damage or disease (Golden et al., 1978) and can indirectly indicate the risk of associated pathologies in dementia evaluations. Significant neurological changes precede cognitive impairments of dementia, and neuropsychological tests included in dementia assessments are often designed to capture the neuropathological processes underlying observable behavioral symptoms sensitively (Fields et al., 2011; Mortamais et al., 2017).

Traditionally, the development of clinical neuropsychological tests has relied on the occurrence and localization of brain lesions as validation criteria. However, advances in neuroimaging and biomarker technologies now allow specific neuropathological features to serve as reference standards (Bilder and Reise, 2019). Neuropathological attributes detected through neuroimaging provide direct and indirect evidence of AD pathology, neurodegeneration in other types of dementia (e.g., frontotemporal dementia), focal brain injuries (e.g., stroke, infarction, traumatic injury), and cerebrovascular disease. This section introduces the characteristics of neuropathologies captured by different neuroimaging methods and their associations with neuropsychological test performance.

Alzheimer's disease pathology

Neuroimaging of AD pathology offers abundant information about the risk of rapidly progressing brain disease, even in cases without clinically observable symptoms. Positron emission tomography (PET) enables the quantification of Alzheimer's-specific pathologies, such as amyloid-beta and tau protein deposits, and reduced glucose metabolism in the temporoparietal regions (Jack et al., 2013; Jagust, 2018). Structural magnetic resonance imaging (MRI) captures the extent to which these pathologies lead to neurodegeneration, such as atrophy in medial temporal lobe structures (Davatzikos et al., 2011; Fjell et al., 2009). These biomarkers often precede clinical diagnoses of dementia by 10–20 years and reveal heterogeneous clusters of dementia subtypes that cannot be easily distinguished based on clinical symptoms alone (Bateman et al., 2012; Dong et al., 2017). While the costliness of biomarkers limits their application in primary care

settings, neuropsychological tests developed using biomarkers as reference criteria can improve diagnostic approaches (Glymour et al., 2018).

Structural and functional measurements of localized brain regions obtained through MRI also serve as neuropathological markers of AD. In the relatively advanced stages of Alzheimer's pathology, structural neurodegeneration observed as global gray matter atrophy provides a prognosis for cognitive decline within a few years (Soldan et al., 2019). Neuropsychological performance reflects functional deficits caused by compromised neural resources that can no longer maintain existing cognitive functions, as well as the demand for additional compensatory neural resources. For example, longitudinal studies tracking brain structure and cognitive performance in older adults diagnosed with Alzheimer's dementia or amnestic MCI over 1–5 years revealed that changes in neuropsychological test scores (e.g., dementia version of the Seoul Neuropsychological Screening Battery) were strongly correlated with cortical volume loss in individuals with MCI (Kim et al., 2017). By contrast, changes in screening tests (e.g., MMSE) only reflected structural changes in individuals already diagnosed with dementia. These findings demonstrate that precise neuropsychological tests, despite their higher time and cost demands, are more sensitive to severe neuropathological changes during early, asymptomatic stages than screening tools.

A systematic review demonstrated that AD pathology, particularly amyloid-beta accumulation, could be detected during preclinical stages, even in the absence of significant impairments in episodic memory, semantic memory, or executive function (Mortamais et al., 2017). Word list learning tasks have been shown to reflect structural changes in the medial temporal lobe, such as hippocampal and entorhinal cortex atrophy, and are strongly associated with amyloid-beta deposition. Poor performance on episodic memory tests corresponds to longitudinal degeneration of temporal cortical and hippocampal structures (Farrell et al., 2018; Fletcher et al., 2018; Hanseeuw et al., 2019).

Several works have validated neuropsychological tests against early-stage Alzheimer's biomarkers. For instance, studies have explored the validity of an interference paradigm in word list learning tasks (where participants learn an interfering list of non-target items) to reflect Alzheimer's pathology (E Curiel, 2013; Loewenstein et al., 2016). This paradigm showed sensitivity to medial temporal lobe volume changes and amyloid-beta accumulation. Composite factor scores combining multiple neuropsychological test measures of executive function have demonstrated stronger correlations with amyloid-beta, tau, and cortical atrophy biomarkers than individual test scores (Gibbons et al., 2012; Gross et al., 2012).

Additionally, composite scores from cognitive tests like word list and story recall, digit-symbol substitution test, and MMSE have been developed to detect cognitive decline in preclinical Alzheimer's stages (Donohue et al., 2014; Hahn et al., 2020; Mormino et al., 2017). Although the precise underlying causes of cognitive deficits can often only be inferred circumstantially, AD pathology appears to reflect not only cortical-level degeneration but also the aggregated effects of white matter microstructural deterioration. This diffuse pathology may explain why tasks that demand the efficiency of inter-module interactions—rather than targeting a specific processing unit—consistently exhibit strong predictive validity for pathological progression. A prime example is the digit-symbol substitution test, which integrates processing speed, executive function, and motor

coordination, thereby effectively capturing the widespread neural network disruptions characteristic of AD (Pichet Binette et al., 2021).

Neurodegeneration

The mechanisms through which dementia-related pathological substances induce neurodegeneration vary between individuals, even among those with the same diagnosis (Noh et al., 2014; Tetreault et al., 2020; Vogel et al., 2021). The prominence of neurodegeneration in specific regions can influence the presentation of dementia subtypes (Tetreault et al., 2020). High-resolution neuroimaging enables the assessment of structural features (e.g., gray matter volume, cortical thickness) and functional attributes (e.g., hemodynamic responses) of affected brain regions. Differences in gray matter volume and functional activation across brain regions indicate variations in information-processing resources. Neuropsychological tests sensitive to regional neurodegeneration provide insights into the severity of neural damage and its impact on specific cognitive processes (Genon et al., 2018, 2022).

The assumption that neuropsychological tests reflect regional neural attributes may align with the traditional neuropsychological principle of double dissociation. According to this principle, selective deficits in specific cognitive tasks resulting from localized brain damage suggest that the tasks assess the unique functions of the localized brain regions (Teuber, 1955; Vaidya et al., 2019). With neuroimaging, this logic generalizes to subtle neuroimaging measures of signal intensity (i.e., voxel intensity) that are not overtly observed as lesions. For example, a series of works showed that long-term retention in word list tasks correlated with entorhinal cortex functionality, while visuospatial pattern recognition tasks correlated with dentate gyrus activity (Brickman et al., 2011, 2014). Similarly, recollection and recognition tasks showed double dissociation of correlating patterns, supporting their validity as measures of distinct medial temporal lobe structures (Argyropoulos et al., 2022; Yonelinas et al., 2022; Yonelinas et al., 2007). Additional research has revealed that learning and retention functions in episodic memory tasks reflect different neural patterns and gray matter volumes (Casaletto et al., 2017; Chang et al., 2010). Such findings underscore the selective sensitivity of neuropsychological measures to specific neural substrates and their utility in detecting diverse dementia pathologies during asymptomatic stages (Whitwell et al., 2009).

White matter pathology and cerebrovascular disease

White matter refers to the neural tracts connecting distant brain regions, supported by microvascular networks that are particularly susceptible to dysfunction and damage (Prins and Scheltens, 2015). This structural characteristic can signify neurocognitive function distinct from gray matter morphology or functional activation. Vascular impairments can lead to damage in adjacent brain structures, a hallmark diagnostic feature of vascular dementia. However, white matter pathology is not exclusive to vascular dementia. Generalized neuropathological changes, including amyloid-beta and tau accumulation (characteristic of AD) and medial temporal lobe atrophy, also reflect potential risks associated with white matter

degeneration (Boyle et al., 2013; Brickman et al., 2009; Lockhart and DeCarli, 2014). Structural brain imaging can identify white matter abnormalities, such as microbleeds or localized infarcts, which appear as hyperintense signals on imaging and provide evidence of the extent of structural neuropathology.

Imaging studies sensitive to white matter lesions have demonstrated that even in clinically asymptomatic older adults, subtle white matter changes and their spatial distribution are associated with increased risks of cognitive impairment. Neuropsychological tests measuring executive function and processing speed are particularly sensitive to white matter integrity and lesion burden (Birdsill et al., 2014; Debette and Markus, 2010; Habes et al., 2018; Hedden et al., 2012). Additionally, the specific location of white matter lesions correlates with declines in episodic memory performance (Brickman et al., 2018; Lockhart et al., 2012; Parks et al., 2011; Smith et al., 2011). Subdividing lesion patterns based on neural tract locations allows researchers to identify whether strategically significant white matter regions contribute to declines in neuropsychological performance (Brugulat-Serrat et al., 2019; Duering et al., 2013; Jiang et al., 2018).

Processing speed tests, rather than working memory tests, are specifically observed to be the most sensitive to cerebrovascular disease (Vasquez and Zakzanis, 2015). Sensitivity to white matter pathology can vary depending on task characteristics, such as whether time constraints are imposed, whether the task requires isolated modules, or whether processing speed tasks incorporate executive components (Lowry et al., 2021; Papp et al., 2014). These differences have led to the development of screening test combinations incorporating processing speed and motor function tasks to enhance diagnostic accuracy (Brookes et al., 2015; Kang et al., 2009).

Another neuropathology associated with dementia is stroke, which includes infarctions, ischemia, and hemorrhages. These events result in invasive brain damage, with behavioral deficits varying according to the locations of the lesions. Lesion mapping methods account for spatial variability in stroke-related damage at the group level (Sperber and Karnath, 2018). By superimposing brain lesion maps of patients exhibiting specific neuropsychological deficits, researchers can identify the brain regions most consistently associated with these impairments. Such analyses have validated neuropsychological tests designed to measure frontal lobe functions and factor-based intelligence scores by linking them to the affected brain regions (Gläscher et al., 2009, 2012). Accumulated lesion-mapping research suggests that highly localized functions, such as visual and motor skills, align closely with cortical regions, whereas tests for complex functions requiring extensive neural collaboration are more indicative of subcortical structures, white matter integrity, and disruptions in functional neural networks (Corbetta et al., 2015; Siegel et al., 2016; Sperber et al., 2022).

Construct validity

In psychometric traditions, construct validation has been emphasized alongside criterion-based approaches (Campbell and Fiske, 1959; Cronbach and Meehl, 1955). Psychological constructs may emerge from social values, linguistic conventions, or practical utility (e.g., composite scores for socioeconomic status). However, from a realist perspective, validation seeks to reflect actual physical

properties beyond the mere utility of the instrument (Fried, 2017). The process of validating a psychological test involves building theoretical foundations that extend beyond its practical application to represent meaningful psychological attributes (Cizek, 2016; Hughes, 2017).

Neuropsychological assessment incorporates cognitive tasks derived from experimental paradigms, applying theories of brain-behavior relationships. While criterion validity focuses on utility (e.g., clinical diagnosis or detection of neuropathology), construct validity requires a focus on ontology (whether the test measures an existing property) and epistemology (whether the property is measurable). Neuropsychological tests, like other psychological tests, must elucidate the cognitive abilities they assess and clarify the attributes contributing to individual differences in performance. Without such clarity, even tests with strong clinical discriminative power may struggle to provide an interpretation of how the prediction occurred (Strauss and Smith, 2009). For instance, MMSE, Montreal Cognitive Assessment (MoCA), and Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog) provide a summary score of multiple item modules that are difficult to apply detailed theoretical interpretation. Put differently, even when an individual is flagged as clinically at risk by an elevated total score, these instruments may lack a structured delineation of subconstructs within cognitive domains; consequently, the fine-grained interpretive meaning can vary with numerous extraneous factors (e.g., relative cognitive strengths, premorbid intellectual profile), making it difficult to attribute the observed global impairment to a specific underlying deficit. Despite the relative lack of evidence, construct validation is essential for explaining phenomena and understanding the cognitive processes underlying clinical outcomes (Jewsbury and Bowden, 2017; Pawlowski et al., 2013). This review explores construct validity by examining two key dimensions: (1) evidence from correlations among test performances (internal structure) and (2) evidence derived from analyses of cognitive processes.

Evidence of internal structure

A common method of establishing construct validity involves examining the nomological network of correlations among variables (Embretson, 1998). This approach assumes that if unobservable attributes (e.g., cognitive constructs) exist, latent variables should causally influence observable behaviors (e.g., test performance), resulting in systematic covariation (Borsboom et al., 2004). Constructs are not directly observable; instead, their theoretical validity is inferred from expected relationships among multiple measures. For example, executive control functions are not represented by a single test but emerge from patterns of correlations across tasks involving different stimuli (e.g., visual, auditory, interfering elements), response modes (e.g., verbal, motor coordination), and task instructions (e.g., response targets, set-shifting). These structures of covariation inform theoretical frameworks about constructs (Friedman and Miyake, 2017; Miyake et al., 2000; Park et al., 2012). Within a covariance model, the construct of executive function can be subdivided into components such as inhibition, set-shifting, and updating, each of which can be measured through specific tasks.

Factor analysis of neuropsychological tests clarifies which measurements belong to specific constructs and the extent to which individual measures represent those constructs. First, a test's representativeness determines whether it validly reflects a given construct (Brown, 2015). This enables researchers to infer which tests should be included in assessments of targeted cognitive domains, such as those used in diagnosing neurocognitive disorders or dementia, and to evaluate how well they represent those domains (Greenaway et al., 2009; Jones et al., 2024; Park et al., 2012). Inadequate representativeness results in low factor loadings and poor model fit, indicating that the hypothesized model structure fails to explain the observed covariance adequately. Comparing the fitness of different models helps determine whether tests broadly encompass constructs or require finer differentiation. Such cumulative evidence has informed diagnostic frameworks that identify behavioral impairments across distinct cognitive domains as defined in the criteria of neurocognitive disorder (Sachdev et al., 2014).

Factor models also guide the development of indices that enable interpretations of multidimensional measures in a neuropsychological battery (Jahng et al., 2015). For instance, the summed scores of weakly related measures form composites representing heterogeneous and broad constructs, whereas measures sharing latent factors yield theoretically coherent interpretations. Episodic memory, a composite cognitive ability, can be represented by indices combining multiple sensory modalities (i.e., auditory, visual) and memory processing stages (i.e., immediate short-term recall, delayed long-term recall) (Drozdick et al., 2011). Patterns of decline observed in these indices help interpret shared attributes of subtests and summarize refined information about constructs, thereby improving predictions of clinical outcomes and dementia pathology (Crane et al., 2012; Gibbons et al., 2012; Gross et al., 2014).

Despite its prevalent use in psychological test validation, expanding the interpretation of the internal structure approach can be challenging in the field of neuropsychological tests. One critical issue occurs in how much a given test can delineate finite units of construct by testing a structural model. When a certain test includes mixed cognitive elements, the measurements may not clearly align with distinct latent variables. For instance, fluency tests can be classified into varying domains, including executive function, processing speed, expressive language function, and semantic memory (Shao et al., 2014; Whiteside et al., 2016). Similarly, attentional tasks may appear to measure a unified cognitive domain but lack sufficient variance to support distinct subdomains empirically (Treviño et al., 2021). A common practice of factor analysis, which typically does not assume cross-loadings across latent constructs, may underspecify these aforementioned cases of multi-component nature. Moreover, the labeling of test scores can be critical in that how tasks are theoretically categorized can significantly alter interpretations of the number of impaired domains observed in examinees. The arbitrary categorization into cognitive domains can make clinical decision-making less reliable, as the severity, breadth, and underlying pathology of cognitive impairment directly affect whether deficits appear in single or multiple cognitive domains (Petersen et al., 2009).

Another issue arises from inconsistencies between traditional constructs and empirical findings, leading to reduced conceptual

clarity. Higher-order factors often override the attributes of individual tests, particularly when analyses rely on normative populations (Bowden, 2004; Delis et al., 2003; Jacobson et al., 2004). For example, systematic variance in healthy populations may not align with expectations in patients with neuropathology, such as the hallmark long-term memory impairments seen in AD. Factor analyses within a normative population may obscure well-known distinctions between short-term and long-term memory structures, underestimating specific deficits caused by neuropathology (Holdnack et al., 2011). Only a few studies have explored measurement invariance across clinical status, which can explore whether latent factors diverge with the progression of neural dysfunction (Hayden et al., 2011).

The fundamental limitation of using internal structure as validity evidence lies in its inability to disentangle causal effects captured by correlations. From a perspective of realism, test scores are valid only if the target construct "exists" and the attributes of that construct causally generate the variance observed in test scores (Borsboom et al., 2004; Hughes, 2017). In some endeavors that suggest the role of working memory capacity in reasoning ability, studies were able to extend correlating evidence of structural equation modeling by experimentally manipulating additional task components (Engle et al., 1999; Harrison et al., 2015). However, correlations observed in cross-sectional data may reflect causal effects originating from a variety of mixed sources and a prolonged timespan other than underlying latent factors. For instance, the correlations between test performances and the effects of latent variables could arise from cumulative interactions associated with long-term cognitive development and aging processes (Kan et al., 2019; Kievit et al., 2017; Van Der Maas et al., 2017). It is important to acknowledge that these mechanisms can impose constraints on the theoretical extrapolation of causal interventions through latent variables. To extend the evidence of construct validity beyond a mere summary of information, it is important to note other alternative mechanisms that form the factor patterns of data.

Evidence of convergent and discriminant validity

Convergent and discriminant validity are critical components of construct validation. These are often demonstrated by examining correlation patterns among theoretically related and unrelated measures. Convergent validity evidence is established when test scores show higher correlations with measures assumed to assess the same construct, while discriminant validity is evidenced by lower correlations with measures assessing different constructs (Campbell and Fiske, 1959; Stern et al., 2014; Westen and Rosenthal, 2003).

Specifically, tests within the same cognitive domain are theoretically expected to show stronger correlations with each other than with tests from other domains. However, the patterns of convergent and discriminant correlations often do not align with the traditional theoretical distinctions developed in neuropsychology based on lesion cases, and in practice, these domain-based correlations often lack clear distinctions (Dodrill, 1999).

One illustrative case is the shared variance between "hold" tests (which are resistant to neuropathology) and "no-hold" tests (which are sensitive to neuropathology), which often complicates interpretations despite their theoretical distinctions (Greenaway et al., 2009). Specifically, verbal comprehension ability is frequently used as a proxy for cognitive reserve. While this ability shows limited changes due to neuropathology and weak causal links to neurological conditions, it still exhibits notable correlations with neuropsychological measures known to be sensitive to pathology, such as delayed word recall (Siedlecki et al., 2009). Although factor structures and relative correlation patterns may be suggestive of distinctions between constructs, such evidence often fails to characterize the conceptual difference fully. This issue is particularly pronounced for tests that reflect both neuropathology sensitivity and premorbid cognitive ability.

The challenge also holds in the examination of predictive neuromarkers of neuropsychological tests. The finding showed that the multivariate model of brain functional connectivities was predictive of the overall domains of the test (Kwak et al., 2021a). In this way, the network-specific contribution to the test performance has been empirically tested. The finding suggests that meaningful brain-behavior predictability is largely comprised of shared intercorrelations and suggests the challenges in proposing discriminant validity. In other words, the brain's functional characteristics predictive of a certain domain of the test are likely to predict all other test scores of other cognitive domains.

The lack of evidence of discriminant validity is a broader issue in psychometrics, where an overemphasis on convergent validity and proposing over-inclusive constructs paradoxically leads to blurred boundaries between constructs (Lilienfeld and Strother, 2020). When a wide range of measures is not available, researchers frequently rely on high correlations with homologous tests to claim convergent validity. This approach often relies on rejecting a less appropriate null hypothesis ("test will show no significant correlation with a target measure"). However, rigorous validation necessitates quantifying and testing whether the test demonstrates significantly larger correlations with theoretically aligned measures than with less related constructs (Westen and Rosenthal, 2003).

Evidence of cognitive process

The foundation of construct validity lies in establishing theories about cognitive processes that underlie test performance (Brewer and Crano, 2014; Embretson, 1998). Experimental paradigms allow researchers to manipulate task stimuli, rules, and responses systematically, predicting behavioral differences and individual variations in cognitive processes. Such paradigms enable the development of tasks that capture specific cognitive impairments, extending theories about how certain pathological conditions affect behavior (Barch, 2005; Festa et al., 2005; Knight and Silverstein, 2001). For example, experimental tasks can measure how cognitive conflicts (e.g., conflicting stimuli and responses) slow reaction times. When systematic increases in interference conditions slow responses, the cognitive processes targeted by the task can be inferred. If individuals with specific neurological conditions exhibit amplified effects under experimental conditions, it suggests deficits in cognitive processes or resource allocation induced by the experimental manipulation.

Although the experimental task itself is rarely included in the clinical assessment battery, process analysis may form the theoretical basis for refining or modifying existing tests (Brickman et al., 2014; Loewenstein et al., 2016). For instance, tests developed through process-oriented approaches can identify the specific stages of language processing or episodic memory affected by brain damage. These qualitative measures not only clarify the nature of brain damage but also reveal disease risks not captured by composite scores (Delis et al., 1987; Goodglass and Kaplan, 1983). For example, differences in how individuals strategically organize information during memory encoding (e.g., semantic clustering) or intrusive errors during recall provide insights into specific deficits and align with neuropathology-based validity evidence (Kirchhoff et al., 2014; Thomas et al., 2020).

However, process-based approaches face challenges when divorced from other validation frameworks. By focusing on within-person variance, they often neglect inter-individual differences, creating a gap in practical utility. That is, systematic variance in behavior due to experimental conditions does not necessarily translate into meaningful individual differences, limiting the approach's applicability for identifying cognitive attributes. This issue is particularly prevalent in tasks requiring complex cognitive integrations. For example, subtraction methods used to isolate specific cognitive processes (e.g., Stroop test: word reading versus color naming) yield stable experimental effects but often exhibit low psychometric reliability (Eisenberg et al., 2019; Enkavi et al., 2017). Measurement error, rather than targeted cognitive attributes, often accounts for most of the observed variance (Hedge et al., 2017).

Disagreement of validity evidence

Disagreements within criterion validity

The diversity of validity evidence extends beyond conceptual differences to include the critical issue of potential inconsistencies among different forms of evidence. For example, individuals with increased measures of cognitive reserve may reach functional thresholds at a more delayed time point in later life. This in turn leads to the observation that tests with concurrent criterion validity in predicting clinical outcomes like functional impairment in dementia (Tucker-Drob, 2019). However, these measures do not directly correspond to the progression of AD pathology. In other words, there is a misalignment between two criterion variables: functional impairment and the presence of neuropathology. Such disagreements are reflected in ongoing debates about the operational definitions of AD as a biological disorder versus a clinical syndrome (Jack et al., 2024; Petersen et al., 2021). As biomarkers become increasingly accessible, the incompatibility between clinical manifestations and biological markers has become more apparent. This necessitates a clearer delineation of the operational criteria that neuropsychological tests are validated against (Hazan et al., 2024).

In the case of the criterion within brain pathology, inconsistencies arise due to differences in the specific characteristics of pathology. For instance, while neuropsychological batteries used in dementia assessments often predict the severity and progression of dementia, their predictive validity for differentiating subtypes (e.g., Alzheimer's vs. vascular dementia) may not align with each other. Certain subtests,

such as constructional praxis, provide strong discriminative information about dementia subtypes but are less sensitive to overall clinical severity (Kwak et al., 2023). These discrepancies may arise from the qualitative differences between focal brain lesions specified from a classical validation design (i.e., clinical case–control) and diffuse disruptions seen in dementia-related neuropathology. Focal lesions, such as those caused by stroke, produce specific behavioral impairments tied to discrete brain regions. In contrast, the distributed nature of neuropathological changes in dementia may limit the utility of narrowly focused cognitive tasks in capturing widespread neural deficits.

The progressive and staged nature of dementia also complicates the validation target. Sensitivity to cognitive domains may vary across disease stages, even within the same dementia type (Kim et al., 2024; Levin et al., 2021). For example, delayed recall functions, which reflect retention processes, often show pronounced declines in later stages of dementia but may not fully capture changes during the preclinical phase. By contrast, executive function tasks such as the Digit Symbol Substitution Test can provide additional information about early-stage changes that memory tasks alone fail to explain.

Disagreements between construct and criterion validation

The mismatch between construct clarity and criterion validity often leads to a tradeoff in the neuropsychological test aim. Construct validity involves refining experimental conditions to isolate specific cognitive processes. For example, tasks may be designed to manipulate stimuli and response rules parametrically or subtract baseline conditions to enhance the precision of targeted constructs. By controlling extraneous variance, test developers aim to hone the clarity of the constructs being measured. However, this refinement may not translate into stronger ecological validity or predictive power for everyday functioning. Paradoxically, tasks designed to measure specific cognitive processes often demonstrate weaker criterion validity for predicting real-life outcomes (Gold et al., 2012; Sbordone, 1997). This has been termed as two approaches of conceptual framework: verisimilitude and veridicality (Chaytor and Schmitter-Edgecombe, 2003). Verisimilitude is the degree to which the cognitive demands of a test theoretically resemble the cognitive demands in the everyday environment, whereas veridicality refers to the degree to which existing tests are empirically related to measures of everyday functioning (Franzen and Wilhelm, 1996).

Specifically, a nuanced tradeoff can be found in the case of cognitive ability tests. Tasks with more clear specification of the cognitive construct may provide less criterion-relevant information. For example, while general cognitive ability strongly predicts clinical diagnoses, specific abilities reflected in each cognitive domain often contribute a limited amount of incremental information (Breit et al., 2024). Likewise, the verbal comprehension index from intelligence tests is reflective of cumulative developmental experiences rather than an explicit sampling of cognitive processes. Despite their construct ambiguity and unspecifiable process, these measures often show robust predictive validity for clinical impairment (Ackerman, 2022; Royall and Palmer, 2012).

This trade-off is also observed in other performance-based tests. The clock drawing test, for instance, requires multiple domains of the

process—executive function, visuospatial construction, motor skills, and semantic knowledge—but inevitably lacks detailed specification of how cognitive processes contribute to the observed functioning. Nonetheless, it remains effective in detecting neurocognitive impairments due to dementia (Tsoi et al., 2015). Similarly, real-life tasks like the Multiple Errands Test (MET) have demonstrated strong psychometric properties but remain uncertain about how specific cognitive domains, such as executive function, constitute the overall performance (Rotenberg et al., 2020).

The challenges in the divergence between construct validity and ecologically oriented criterion validity may become more prominent in tasks of complexity and multiplicity of components. It is more difficult to standardize and control experimental structure in tasks requiring a certain time for reasoning and mobilizing strategies. Some of the works have designed ways to assess by determining the major everyday cognitive ability required by the environment (Chaytor and Schmitter-Edgecombe, 2003). High-order cognitive abilities, however, often involve heterogeneous strategies and functional brain activity patterns, even among individuals with similar test scores (Barulli et al., 2013; Seghier and Price, 2018). This tradeoff reflects the adaptive and flexible nature of real-life problem-solving, which depends on the dynamic integration of cognitive processes rather than the efficiency of a single isolated process. Consequently, it would be less surprising to find that cognitive tasks emphasizing ecological validity, even at the expense of construct refinement, may hold greater utility in dementia evaluations (Chaytor and Schmitter-Edgecombe, 2003; García-Betances et al., 2015; Howett et al., 2019; Kessels, 2019).

Choices of scoring practices also regard the tradeoff between clarity of construct and predictability of criterion. Neuropsychological tests typically exhibit positive manifolds (i.e., intercorrelations) among subtests, enabling the aggregation of composite scores based on shared variance (Agelink van Rentergem et al., 2020). A total score across multiple cognitive domains or summing subtests as a composite score often shows robust predictive validity for clinical criteria, providing sensitive and accurate information about impairments (Chandler et al., 2005; Fields et al., 2010; Wolfsgruber et al., 2014). However, this expansion in criterion validity may come at the expense of construct clarity. Neuropsychological batteries tend to have lower internal consistency than factor-based ability tests, such as intelligence tests, indicating limited homogeneity of constructs underlying total scores (Jones et al., 2024; Kiselica et al., 2020). This suggests that improved criterion validity may be partially achieved by compromising the conceptual coherence of constructs.

Digital markers: redefining validation standards

Traditional neuropsychological assessments have relied on a concise set of scores to evaluate predictive validity and construct-based interpretations. However, advances in digital markers introduce challenges that require alternative validation approaches (Kumar et al., 2021). First, machine learning (ML) algorithms no longer rely on the linear combinations of individual scores. For instance, adding subtests in a battery may fail to improve criterion validity under linear models, whereas such addition yields significant gains in predictive accuracy with nonlinear ML models (Kwak et al., 2022). Such superiority of nonlinear algorithms is well-documented across various predictive

domains (Schulz et al., 2020). This indicates that the validity of individual test scores contributing to diagnostic information can vary depending on the specific ML algorithm employed.

Predictive modeling becomes increasingly critical as the dimensionality of available information grows (Yarkoni and Westfall, 2017). Digital phenotyping involves high-dimensional data from device logs, daily activity metrics, wearable sensors, and voice features (Palliya Guruge et al., 2021). Although the digital data typically lacks simple summary measures, they provide sparsely distributed predictive information across multiple features which overall contributes to approximating neurocognitive function to some extent (Hackett and Giovannetti, 2022; Harris et al., 2024; Holmes et al., 2022). From the perspective of validity evidence, a principal challenge is that these approaches yield such a profusion of variables that individual verification becomes infeasible, thereby complicating construct-based interpretations of criterion validity. Moreover, in such contexts, the predictive power of individual predictor variables and the utility of the measurement modality may lead to divergent inferences. For example, even if dozens of voice features exhibit poor criterion validity and one of numerous line-drawing features demonstrate excellent criterion validity, the feature set may nevertheless achieve superior aggregate criterion validity when the assessment module is evaluated at the whole set level. A "simulated lesion" analytic strategy—designed to assess the overall contribution of a specific modality set—can quantify both how many assessment modules (e.g., voice and drawing module) should be included and the magnitude of their practical usefulness with respect to the criterion (Hahn et al., 2017; Kohoutová et al., 2020).

Conclusion

This review examined the contemporary issues underlying the criteria for deeming neuropsychological tests as "valid" for dementia assessment. It also explored how advances in neuroimaging and biomarker research necessitate changes from traditional validation processes. Unlike the conventional approach of validating the correspondence between constructs and criteria through studies on clinical cases of brain damage, the proliferation of biomarkers and neuroimaging technologies—capable of identifying diverse attributes of brain pathology (e.g., regional patterns, functional deficits, and types of pathology)—has broadened the validity criteria supporting the utility of neuropsychological tests. Additionally, this review underscored the need to re-examine the varied and assumption-specific nature of evidence for validity.

It is emphasized that neuropsychological tests may fail to meet various validity criteria in a consistent manner, not necessarily due to their inherent limitations of validity, but potentially because of fundamental discordance of the conceptual status of validity. This discussion highlights that validity is not monolithic and that different tools may meet distinct criteria depending on their specific constructs and intended applications.

The accompanying figure illustrates the conceptual framework within which the validity of neuropsychological tests for dementia evaluation is established. Depending on the test, the domain of evidence may vary, influenced by the relative strength of construct and
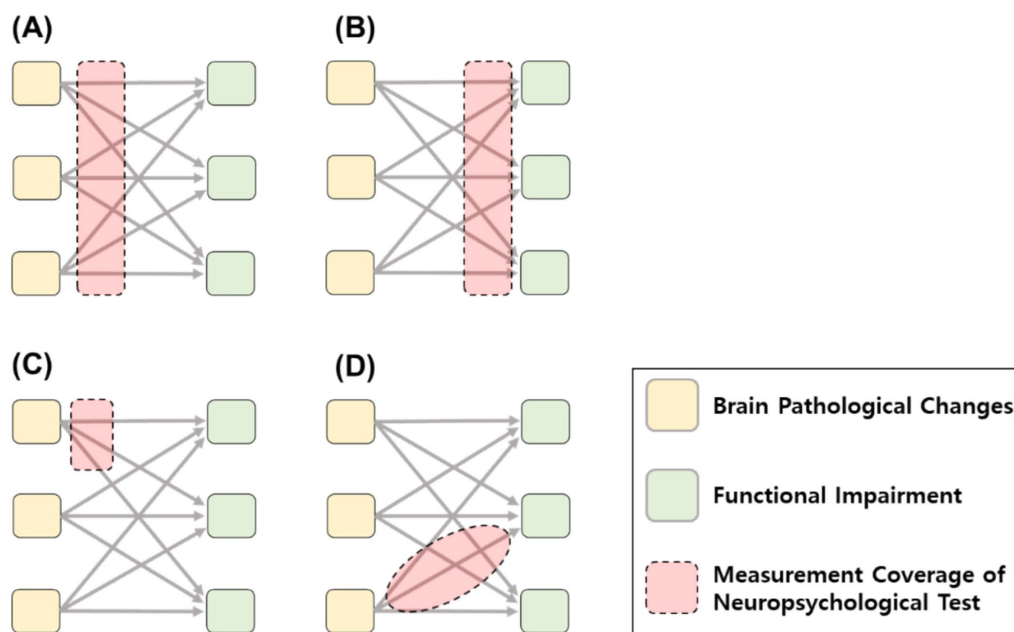
FIGURE 1
Various validity positions and goals of measurement coverage of neuropsychological test. **(A)** Measurement coverage sensitive to brain pathology criterion. **(B)** Measurement coverage sensitive to functional impairment criterion. **(C)** Measurement coverage to specific neuropathology changes. **(D)** Measurement coverage corresponding to specific construct.

criterion validity, as well as the scope and specificity of measurement. For instance, tests may differ in their measurement emphasis on either sensitivity to pathology underlying the disease (Figure 1A), ecological validity and sensitivity to functional outcomes (Figure 1B), capacity to characterize unique attributes of brain pathology (Figure 1C), or their ability to capture specific processing units and explain the continuum from disease to disorder with robust construct validity (Figure 1D).

This review also invites a more refined generalization of the evidence-based assessment approach into the practice of neuropsychological test validation (Hunsley and Mash, 2007). However, some neuropsychological test cases can be challenged by fundamentally disagreeing validity evidence and trade-offs between the aims of evidence, where the primacy of one type of evidence may compromise another. In other words, the validity of neuropsychological tools needs to be evaluated under their specific purposes and contexts, rather than seeking straightforward alignment of validity evidence as a parallel criterion.

Ultimately, validity evidence reflects not only the inherent soundness of an instrument's internal logic but also its dependence on external contexts. Researchers and practitioners are encouraged to explicitly articulate the specific contexts in which validity evidence is established, enabling more precise interpretations of such evidence. Conversely, usage guidelines can be formulated under context-specific validity constraints—for example, by explicitly noting that validity has not yet been demonstrated in certain contexts, or by directing that interpretation be expressly deferred for particular subconstructs within a cognitive domain. This contextualized perspective of the validation process will be critical for advancing the clarity and utility of validity claims in neuropsychological assessment.

Author contributions

SK: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

Funding

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author declares that Gen AI was used in the creation of this manuscript. Chat GPT-4o was used for the improvement in readership and refinement of the writing.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ackerman, P. L. (2022). Intelligence process vs. content and academic performance: a trip through a house of mirrors. *J. Intelligence* 10:128. doi: 10.3390/jintelligence10040128

Agelink van Rentergem, J. A., de Vent, N. R., Schmand, B. A., Murre, J. M. J., Staaks, J. P. C., and Huizenga, H. M. (2020). The factor structure of cognitive functioning in cognitively healthy participants: a meta-analysis and meta-analysis of individual participant data. *Neuropsychol. Rev.* 30, 51–96. doi: 10.1007/s11065-019-09423-6

American Educational Research Association (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Amieva, H., Mokri, H., Le Goff, M., Meillon, C., Jacqmin-Gadda, H., Foubert-Samier, A., et al. (2014). Compensatory mechanisms in higher-educated subjects with Alzheimer's disease: a study of 20 years of cognitive decline. *Brain* 137, 1167–1175. doi: 10.1093/brain/awu035

Anastasi, A. (1950). The concept of validity in the interpretation of test scores. *Educ. Psychol. Meas.* 10, 67–78. doi: 10.1177/001316445001000105

Argyropoulos, G. P. D., Dell'Acqua, C., Butler, E., Loane, C., Roca-Fernandez, A., Almozel, A., et al. (2022). Functional specialization of the medial temporal lobes in human recognition memory: dissociating effects of hippocampal versus Parahippocampal damage. *Cereb. Cortex* 32, 1637–1652. doi: 10.1093/cercor/bhab290

Bäckman, L., Jones, S., Berger, A. K., Laukka, E. J., and Small, B. J. (2005). Cognitive impairment in preclinical Alzheimer's disease: a meta-analysis. *Neuropsychology* 19, 520–531. doi: 10.1037/0894-4105.19.4.520

Barch, D. M. (2005). The cognitive Neuroscience of schizophrenia. *Annu. Rev. Clin. Psychol.* 1, 321–353. doi: 10.1146/annurev.clinpsy.1.102803.143959

Barulli, D. J., Rakitin, B. C., Lemaire, P., and Stern, Y. (2013). The influence of cognitive reserve on strategy selection in normal aging. *J. Int. Neuropsychol. Soc.* 19, 841–844. doi: 10.1017/S1355617713000593

Bateman, R. J., Xiong, C., Benzinger, T. L. S. S., Fagan, A. M., Goate, A., Fox, N. C., et al. (2012). Clinical and biomarker changes in dominantly inherited Alzheimer's disease. *N. Engl. J. Med.* 367, 795–804. doi: 10.1056/NEJMoa1202753

Belleville, S., Fouquet, C., Duchesne, S., Collins, D. L., and Hudon, C. (2014). Detecting early preclinical Alzheimer's disease via cognition, neuropsychiatry, and neuroimaging: qualitative review and recommendations for testing. *J. Alzheimer's Dis.* 42, S375–S382. doi: 10.3233/JAD-141470

Bilder, R. M. (2011). Neuropsychology 3.0: evidence-based science and practice. *J. Int. Neuropsychol. Soc.* 17, 7–13. doi: 10.1017/S1355617710001396

Bilder, R. M., and Reise, S. P. (2019). Neuropsychological tests of the future: how do we get there from here? *Clin. Neuropsychol.* 33, 220–245. doi: 10.1080/13854046.2018.1521993

Bilgel, M., An, Y., Lang, A., Prince, J., Ferrucci, L., Jedynak, B., et al. (2014). Trajectories of Alzheimer disease-related cognitive measures in a longitudinal sample. *Alzheimers Dement.* 10, 735–742.e4. doi: 10.1016/j.jalz.2014.04.520

Birdsill, A. C., Koscik, R. L., Jonaitis, E. M., Johnson, S. C., Okonkwo, O. C., Hermann, B. P., et al. (2014). Regional white matter hyperintensities: aging, Alzheimer's disease risk, and cognitive function. *Neurobiol. Aging* 35, 769–776. doi: 10.1016/j.neurobiolaging.2013.10.072

Borsboom, D., Mellenbergh, G. J., and van Heerden, J. (2004). The concept of validity. *Psychol. Rev.* 111, 1061–1071. doi: 10.1037/0033-295X.111.4.1061

Bowden, S. C. (2004). The role of factor analysis in construct validity: is it a myth? *J. Int. Neuropsychol. Soc.* 10, 1018–1019. doi: 10.1017/S1355617704107091

Boyle, P. A., Wilson, R. S., Yu, L., Barr, A. M., Honer, W. G., Schneider, J. A., et al. (2013). Much of late life cognitive decline is not due to common neurodegenerative pathologies. *Ann. Neurol.* 74, 478–489. doi: 10.1002/ana.23964

Breit, M., Scherrer, V., and Preckel, F. (2024). How useful are specific cognitive ability scores? An investigation of their stability and incremental validity beyond general intelligence. *Intelligence* 103:101816. doi: 10.1016/j.intell.2024.101816

Brewer, M. B., and Crano, W. D. (2014). "Research design and issues of validity" in Handbook of research methods in social and personality psychology. New York, NY, USA: Cambridge University Press.

Brickman, A. M., Khan, U. A., Provenzano, F. A., Yeung, L.-K., Suzuki, W., Schroeter, H., et al. (2014). Enhancing dentate gyrus function with dietary flavanols improves cognition in older adults. *Nat. Neurosci.* 17, 1798–1803. doi: 10.1038/nn.3850

Brickman, A. M., Muraskin, J., and Zimmerman, M. E. (2009). Structural neuroimaging in Altheimer's disease: do white matter hyperintensities matter? *Dialogues Clin. Neurosci.* 11, 181–190. doi: 10.1016/j.siny.2015.10.004

Brickman, A. M., Stern, Y., and Small, S. A. (2011). Hippocampal subregions differentially associate with standardized memory tests. *Hippocampus* 21, 923–928. doi: 10.1002/hipo.20840

Brickman, A. M., Tosto, G., Gutierrez, J., Andrews, H., Gu, Y., Narkhede, A., et al. (2018). An MRI measure of degenerative and cerebrovascular pathology in Alzheimer disease. *Neurology* 91, e1402–e1412. doi: 10.1212/WNL.0000000000006310

Brookes, R. L., Hollocks, M. J., Khan, U., Morris, R. G., and Markus, H. S. (2015). The brief memory and executive test (BMET) for detecting vascular cognitive impairment in small vessel disease: a validation study. *BMC Med.* 13, 1–8. doi: 10.1186/s12916-015-0290-y

Brown, T. A. (2015). Confirmatory factor analysis for applied research. *2nd* Edn. New York: Guilford Press.

Brugulat-Serrat, A., Salvadó, G., Sudre, C. H., Grau-Rivera, O., Suárez-Calvet, M., Falcon, C., et al. (2019). Patterns of white matter hyperintensities associated with cognition in middle-aged cognitively healthy individuals. *Brain Imaging Behav.* 14, 2012–2023. doi: 10.1007/s11682-019-00151-2

Campbell, D. T., and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* 56, 81–105. doi: 10.1037/H0046016

Casaletto, K. B., Marx, G., Dutt, S., Neuhaus, J., Saloner, R., Kritikos, L., et al. (2017). Is "learning" episodic memory? Distinct cognitive and neuroanatomic correlates of immediate recall during learning trials in neurologically normal aging and neurodegenerative cohorts. *Neuropsychologia* 102, 19–28. doi: 10.1016/j.neuropsychologia.2017.05.021

Chandler, M. J., Lacritz, L. H., Hynan, L. S., Barnard, H. D., Allen, G., Deschner, M., et al. (2005). A total score for the CERAD neuropsychological battery. *Neurology* 65, 102–106. doi: 10.1212/01.wnl.0000167607.63000.38

Chang, Y. L., Bondi, M. W., Fennema-Notestine, C., McEvoy, L. K., Hagler, D. J., Jacobson, M. W., et al. (2010). Brain substrates of learning and retention in mild cognitive impairment diagnosis and progression to Alzheimer's disease. *Neuropsychologia* 48, 1237–1247. doi: 10.1016/j.neuropsychologia.2009.12.024

Chapman, R. M., Mapstone, M., McCrary, J. W., Gardner, M. N., Porsteinsson, A., Sandoval, T. C., et al. (2011). Predicting conversion from mild cognitive impairment to Alzheimer's disease using neuropsychological tests and multivariate methods. *J. Clin. Exp. Neuropsychol.* 33, 187–199. doi: 10.1080/13803395.2010.499356

Chaytor, N., and Schmitter-Edgecombe, M. (2003). The ecological validity of neuropsychological tests: a review of the literature on everyday cognitive skills. *Neuropsychol. Rev.* 13, 181–197. doi: 10.1023/B:NERV.0000009483.91468.fb

Chen, J. (2018). A hierarchical taxonomy of test validity for more flexibility of validation. *Front. Psychol.* 9:972. doi: 10.3389/fpsyg.2018.00972

Cizek, G. J. (2016). Validating test score meaning and defending test score use: different aims, different methods. *Assess. Educ. Princ. Policy Pract.* 23, 212–225. doi: 10.1080/0969594X.2015.1063479

Corbetta, M., Ramsey, L., Callejas, A., Baldassarre, A., Hacker, C. D., Siegel, J. S., et al. (2015). Common behavioral clusters and subcortical anatomy in stroke. *Neuron* 85, 927–941. doi: 10.1016/j.neuron.2015.02.027

Crane, P. K., Carle, A., Gibbons, L. E., Insel, P., Mackin, R. S., Gross, A., et al. (2012). Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Brain Imaging Behav.* 6, 502–516. doi: 10.1007/s11682-012-9186-z

Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychol. Bull.* 52, 281–302. doi: 10.1037/h0040957

Davatzikos, C., Bhatt, P., Shaw, L. M., Batmanghelich, K. N., and Trojanowski, J. Q. (2011). Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiol. Aging* 32, 2322.e19–2322.e27. doi: 10.1016/j.neurobiolaging.2010.05.023

Debette, S., and Markus, H. S. (2010). The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. *BMJ* 341:c3666. doi: 10.1136/bmj.c3666

Delis, D. C., Jacobson, M., Bondi, M. W., Hamilton, J. M., and Salmon, D. P. (2003). The myth of testing construct validity using factor analysis or correlations with normal or mixed clinical populations: lessons from memory assessment. *J. Int. Neuropsychol. Soc.* 9, 936–946. doi: 10.1017/S1355617703960139

Delis, D. C., Kramer, J. H., Kaplan, E., and Thompkins, B. A. O. (1987). CVLT: California verbal learning test-adult version: manual: Psychological Corporation.

Demakis, G. J. (2004). Frontal lobe damage and tests of executive processing: a meta-analysis of the category test, Stroop test, and trail-making test. *J. Clin. Exp. Neuropsychol.* 26, 441–450. doi: 10.1080/13803390490510149

Dodrill, C. B. (1999). Myths of neuropsychology: further considerations. *Clin. Neuropsychol.* 13, 562–572. doi: 10.1076/1385-4046(199911)13:04;1-Y;FT562

Donders, J. (2020). The incremental value of neuropsychological assessment: a critical review. *Clin. Neuropsychol.* 34, 56–87. doi: 10.1080/13854046.2019.1575471

Dong, A., Toledo, J. B., Honnorat, N., Doshi, J., Varol, E., Sotiras, A., et al. (2017). Heterogeneity of neuroanatomical patterns in prodromal Alzheimer's disease: links to cognition, progression and biomarkers. *Brain* 145, 732–747. doi: 10.1093/brain/aww319

Donohue, M. C., Sperling, R. A., Salmon, D. P., Rentz, D. M., Raman, R., Thomas, R. G., et al. (2014). The preclinical Alzheimer cognitive composite: measuring amyloid-related decline. *JAMA Neurol.* 71, 961–970. doi: 10.1001/JAMANEUROL.2014.803

Drozdick, L. W., Holdnack, J. A., and Hilsabeck, R. C. (2011). "Essentials of WMS®-IV assessment" in Essentials of WMS®-IV assessment. Hoboken, NJ, USA: Wiley.

Dubois, B., Slachevsky, A., Litvan, I., and Pillon, B. (2000). The FAB: a frontal assessment battery at bedside. *Neurology* 55, 1621–1626. doi: 10.1212/WNL.55.11.1621

Duering, M., Gonik, M., Malik, R., Zieren, N., Reyes, S., Jouvent, E., et al. (2013). Identification of a strategic brain network underlying processing speed deficits in vascular cognitive impairment. *NeuroImage* 66, 177–183. doi: 10.1016/j.neuroimage.2012.10.084

E Curiel, R. (2013). A new scale for the evaluation of proactive and retroactive interference in mild cognitive impairment and early Alzheimer's disease. *J. Aging Sci.* 1, 1–5. doi: 10.4172/2329-8847.1000102

Eisenberg, I. W., Bissett, P. G., Zeynep Enkavi, A., Li, J., MacKinnon, D. P., Marsch, L. A., et al. (2019). Uncovering mental structure through data-driven ontology discovery. *Nature Commun.* 10:2319. doi: 10.1038/s41467-019-10301-1

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: application to abstract reasoning. *Psychol. Methods* 3, 380–396. doi: 10.1037/1082-989X.3.3.380

Engle, R. W., Tuholski, S. W., Laughlin, J. E., and Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *J. Exp. Psychol. Gen.* 128, 309–331. doi: 10.1037/0096-3445.128.3.309

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., David, P., Marsch, L. A., et al. (2017). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Procee National Acad Sci.* 116, 5472–5477. doi: 10.1073/pnas.1818430116

Farias, S. T., Harrell, E., Neumann, C., and Houtz, A. (2003). The relationship between neuropsychological performance and daily functioning in individuals with Alzheimer's disease: ecological validity of neuropsychological tests. *Arch. Clin. Neuropsychol.* 18, 655–672. doi: 10.1016/S0887-6177(02)00159-2

Farrell, M. E., Chen, X., Rundle, M. M., Chan, M. Y., Wig, G. S., and Park, D. C. (2018). Regional amyloid accumulation and cognitive decline in initially amyloid-negative adults. *Neurology* 91, e1809–e1821. doi: 10.1212/WNL.0000000000006469

Festa, E. K., Insler, R. Z., Salmon, D. P., Paxton, J., Hamilton, J. M., and Heindel, W. C. (2005). Neocortical disconnectivity disrupts sensory integration in Alzheimer's disease. *Neuropsychology* 19, 728–738. doi: 10.1037/0894-4105.19.6.728

Fields, J. A., Ferman, T. J., Boeve, B. F., and Smith, G. E. (2011). Neuropsychological assessment of patients with dementing illness. *Nat. Rev. Neurol.* 7, 677–687. doi: 10.1038/nrneurol.2011.173

Fields, J. A., Machulda, M., Aakre, J., Ivnik, R. J., Boeve, B. F., Knopman, D. S., et al. (2010). Utility of the DRS for predicting problems in day-to-day functioning. *Clin. Neuropsychol.* 24, 1167–1180. doi: 10.1080/13854046.2010.514865

Fjell, A. M., Walhovd, K. B., Fennema-Notestine, C., McEvoy, L. K., Hagler, D. J., Holland, D., et al. (2009). One-year brain atrophy evident in healthy aging. *J. Neurosci.* 29, 15223–15231. doi: 10.1523/JNEUROSCI.3252-09.2009

Fletcher, E., Filshtein, T. J., Harvey, D., Renaud, A., Mungas, D., and DeCarli, C. (2018). Staging of amyloid β, t-tau, regional atrophy rates, and cognitive change in a nondemented cohort: results of serial mediation analyses. *Alzheimers Dement. (N Y)* 10, 382–393. doi: 10.1016/j.dadm.2018.04.001

Franzen, M. D., and Wilhelm, K. L. (1996). "Conceptual foundations of ecological validity in neuropsychological assessment" in Ecological validity of neuropsychological testing. (Delray Beach, FL: Gr Press/St Lucie Press, Inc), 91–112.

Fried, E. I. (2017). What are psychological constructs? On the nature and statistical modelling of emotions, intelligence, personality traits and mental disorders. *Health Psychol. Rev.* 11, 130–134. doi: 10.1080/17437199.2017.1306718

Friedman, N. P., and Miyake, A. (2017). Unity and diversity of executive functions: individual differences as a window on cognitive structure. *Cortex* 86, 186–204. doi: 10.1016/J.CORTEX.2016.04.023

García-Betances, R. I., Arredondo Waldmeyer, M. T., Fico, G., and Cabrera-Umpiérrez, M. F. (2015). A succinct overview of virtual reality technology use in Alzheimer's disease. *Front. Aging Neurosci.* 7:80. doi: 10.3389/fnagi.2015.00080

Genon, S., Eickhoff, S. B., and Kharabian, S. (2022). Linking interindividual variability in brain structure to behaviour. *Nat. Rev. Neurosci.* 23, 307–318. doi: 10.1038/s41583-022-00584-7

Genon, S., Reid, A., Langner, R., Amunts, K., and Eickhoff, S. B. (2018). How to characterize the function of a brain region. *Trends Cogn. Sci.* 22, 350–364. doi: 10.1016/j.tics.2018.01.010

Gibbons, L. E., Carle, A. C., Mackin, R. S., Harvey, D., Mukherjee, S., Insel, P., et al. (2012). A composite score for executive functioning, validated in Alzheimer's Disease Neuroimaging Initiative (ADNI) participants with baseline mild cognitive impairment. *Brain Imaging Behav.* 6, 517–527. doi: 10.1007/s11682-012-9176-1

Gläscher, J., Adolphs, R., Damasio, H., Bechara, A., Rudrauf, D., Calamia, M., et al. (2012). Lesion mapping of cognitive control and value-based decision making in the prefrontal cortex. *Proc. Natl. Acad. Sci. USA* 109, 14681–14686. doi: 10.1073/pnas.1206608109

Gläscher, J., Tranel, D., Paul, L. K., Rudrauf, D., Rorden, C., Hornaday, A., et al. (2009). Lesion mapping of cognitive abilities linked to intelligence. *Neuron* 61, 681–691. doi: 10.1016/j.neuron.2009.01.026

Glymour, M. M., Brickman, A. M., Kivimaki, M., Mayeda, E. R., Chêne, G., Dufouil, C., et al. (2018). Will biomarker-based diagnosis of Alzheimer's disease maximize scientific progress? Evaluating proposed diagnostic criteria. *Eur. J. Epidemiol.* 33, 607–612. doi: 10.1007/s10654-018-0418-4

Gold, J. M., Barch, D. M., Carter, C. S., Dakin, S., Luck, S. J., MacDonald, A. W., et al. (2012). Clinical, functional, and intertask correlations of measures developed by the cognitive neuroscience test reliability and clinical applications for schizophrenia consortium. *Schizophr. Bull.* 38, 144–152. doi: 10.1093/schbul/sbr142

Golden, C. J., Hammeke, T. A., and Purisch, A. D. (1978). Diagnostic validity of a standardized neuropsychological battery derived from Luria's neuropsychological tests. *J. Consult. Clin. Psychol.* 46, 1258–1265. doi: 10.1037/0022-006X.46.6.1258

Goodglass, H., and Kaplan, E. (1983). "The assessment of aphasia and related disorders" in Boston diagnostic aphasia examination booklet. 2nd ed. eds. Lea and Febiger. Philadelphia, PA, USA.

Greenaway, M. C., Smith, G. E., Tangalos, E. G., Geda, Y. E., and Ivnik, R. J. (2009). Mayo older Americans normative studies: factor analysis of an expanded neuropsychological battery. *Clin. Neuropsychol.* 23, 7–20. doi: 10.1080/13854040801891686

Gross, A. L., Manly, J. J., Pa, J., Johnson, J. K., Park, L. Q., Mitchell, M. B., et al. (2012). Cortical signatures of cognition and their relationship to Alzheimer's disease. *Brain Imaging Behav.* 6, 584–598. doi: 10.1007/s11682-012-9180-5

Gross, A. L., Sherva, R., Mukherjee, S., Newhouse, S., Kauwe, J. S. K., Munsie, L. M., et al. (2014). Calibrating longitudinal cognition in Alzheimer's disease across diverse test batteries and datasets. *Neuroepidemiology* 43, 194–205. doi: 10.1159/000367970

Habes, M., Sotiras, A., Erus, G., Toledo, J. B., Janowitz, D., Wolk, D. A., et al. (2018). White matter lesions spatial heterogeneity, links to risk factors, cognition, genetics, and atrophy. *Neurology* 91, e964–e975. doi: 10.1212/WNL.0000000000006116

Hackett, K., and Giovannetti, T. (2022). Capturing cognitive aging in vivo: application of a neuropsychological framework for emerging digital tools. *JMIR Aging* 5:e38130. doi: 10.2196/38130

Hahn, A., Kim, Y. J., Kim, H. J., Jang, H., Cho, H., Choi, S. H., et al. (2020). The preclinical amyloid sensitive composite to determine subtle cognitive differences in preclinical Alzheimer's disease. *Sci. Rep.* 10:13583. doi: 10.1038/s41598-020-70386-3

Hahn, T., Nierenberg, A. A., and Whitfield-Gabrieli, S. (2017). Predictive analytics in mental health: applications, guidelines, challenges and perspectives. *Mol. Psychiatry* 22, 37–43. doi: 10.1038/mp.2016.201

Hanseeuw, B. J., Betensky, R. A., Jacobs, H. I. L., Schultz, A. P., Sepulcre, J., Becker, J. A., et al. (2019). Association of amyloid and tau with cognition in preclinical Alzheimer disease: a longitudinal study. *JAMA Neurol.* 76, 915–924. doi: 10.1001/jamaneurol.2019.1424

Harris, C., Tang, Y., Birnbaum, E., Cherian, C., Mendhe, D., and Chen, M. H. (2024). Digital neuropsychology beyond computerized cognitive assessment: applications of novel digital technologies. *Arch. Clin. Neuropsychol.* 39, 290–304. doi: 10.1093/arclin/acae016

Harrison, T. L., Shipstead, Z., and Engle, R. W. (2015). Why is working memory capacity related to matrix reasoning tasks? *Mem. Cogn.* 43, 389–396. doi: 10.3758/s13421-014-0473-3

Hayden, K. M., Jones, R. N., Zimmer, C., Plassman, B. L., Browndyke, J. N., Pieper, C., et al. (2011). Factor structure of the national Alzheimer's coordinating centers dataset neuropsychological battery: an evaluation of invariance between and within groups over time. *Alzheimer Dis. Assoc. Disord.* 25, 128–137. doi: 10.1097/WAD.0b013e3181ffa76d

Hazan, J., Liu, K. Y., Costello, H., Isaacs, J. D., Thambisetty, M., and Howard, R. (2024). Challenges in a biological definition of Alzheimer disease. *Neurology* 103:e209884. doi: 10.1212/WNL.0000000000209884

Hedden, T., Mormino, E. C., Amariglio, R. E., Younger, A. P., Schultz, A. P., Becker, J. A., et al. (2012). Cognitive profile of amyloid burden and white matter hyperintensities in cognitively normal older adults. *J. Neurosci.* 32, 16233–16242. doi: 10.1523/JNEUROSCI.2462-12.2012

Hedge, C., Powell, G., and Sumner, P. (2017). The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* 50, 1166–1186. doi: 10.3758/s13428-017-0935-1

Holdnack, J. A., Zhou, X., Larrabee, G. J., Millis, S. R., and Salthouse, T. A. (2011). Confirmatory factor analysis of the WAIS-IV/WMS-IV. *Assessment* 18, 178–191. doi: 10.1177/1073191110393106

Holmes, A. A., Tripathi, S., Katz, E., Mondesire-Crump, I., Mahajan, R., Ritter, A., et al. (2022). A novel framework to estimate cognitive impairment via finger interaction with digital devices. *Brain Commun.* 4:fcac194. doi: 10.1093/braincomms/fcac194

Howett, D., Castegnaro, A., Krzywicka, K., Hagman, J., Marchment, D., Henson, R., et al. (2019). Differentiation of mild cognitive impairment using an entorhinal cortex-based test of virtual reality navigation. *Brain* 142, 1751–1766. doi: 10.1093/brain/awz116

Hughes, D. J. (2017). "Psychometric validity: establishing the accuracy and appropriateness of psychometric measures" in The Wiley handbook of psychometric testing: a multidisciplinary reference on survey, scale and test development. Chichester, West Sussex, UK: Wiley-Blackwell.

Hunsley, J., and Mash, E. J. (2007). Evidence-based assessment. *Annu. Rev. Clin. Psychol.* 3, 29–51. doi: 10.1146/annurev.clinpsy.3.022806.091419

Jack, C. R., Andrews, J. S., Beach, T. G., Buracchio, T., Dunn, B., Graf, A., et al. (2024). Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's association workgroup. *Alzheimers Dement.* 20, 5143–5169. doi: 10.1002/alz.13859

Jack, C. R., Knopman, D. S., Jagust, W. J., Petersen, R. C., Weiner, M. W., Aisen, P. S., et al. (2013). Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *Lancet Neurol.* 12, 207–216. doi: 10.1016/S1474-4422(12)70291-0

Jacobson, M. W., Delis, D. C., Hamilton, J. M., Bondi, M. W., and Salmon, D. P. (2004). How do neuropsychologist define cognitive constructs? Further thoughts on limitations of factor analysis used with normal or mixed clinical populations. *J. Int. Neuropsychol. Soc.* 10, 1020–1021. doi: 10.1017/S1355617704107121

Jagust, W. (2018). Imaging the evolution and pathophysiology of Alzheimer disease. *Nat. Rev. Neurosci.* 19, 687–700. doi: 10.1038/s41583-018-0067-3

Jahng, S., Na, D. L., and Kang, Y. (2015). Constructing a composite score for the Seoul neuropsychological screening battery-core. *Dement. Neurocogn. Disord.* 14:137. doi: 10.12779/dnd.2015.14.4.137

Jang, H., Ye, B. S., Woo, S., Kim, S. W., Chin, J., Choi, S. H., et al. (2017). Prediction model of conversion to dementia risk in subjects with amnestic mild cognitive impairment: a longitudinal, multi-center clinic-based study. *J. Alzheimer's Dis.* 60, 1579–1587. doi: 10.3233/JAD-170507

Jefferson, A. L., Paul, R. H., Ozonoff, A., and Cohen, R. A. (2006). Evaluating elements of executive functioning as predictors of instrumental activities of daily living (IADLs). *Arch. Clin. Neuropsychol.* 21, 311–320. doi: 10.1016/j.acn.2006.03.007

Jessen, F., Amariglio, R. E., van Boxtel, M., Breteler, M., Ceccaldi, M., Chételat, G., et al. (2014). A conceptual framework for research on subjective cognitive decline in preclinical Alzheimer's disease. *Alzheimers Dement.* 10, 844–852. doi: 10.1016/j.jalz.2014.01.001

Jewsbury, P. A., and Bowden, S. C. (2017). "Construct validity has a critical role in evidence-based neuropsychological assessment" in Neuropsychological assessment in the age of evidence-based practice: diagnostic and treatment evaluations. eds. P. A. Jewsbury and S. C. Bowden. (New York, NY: Oxford University Press), 33–63.

Jiang, J., Paradise, M., Liu, T., Armstrong, N. J., Zhu, W., Kochan, N. A., et al. (2018). The association of regional white matter lesions with cognition in a community-based cohort of older individuals. *NeuroImage Clin.* 19, 14–21. doi: 10.1016/j.nicl.2018.03.035

Jones, R. N., Manly, J. J., Langa, K. M., Ryan, L. H., Levine, D. A., McCammon, R., et al. (2024). Factor structure of the harmonized cognitive assessment protocol neuropsychological battery in the health and retirement study. *J. Int. Neuropsychol. Soc.* 30, 47–55. doi: 10.1017/S135561772300019X

Kan, K., Van Der Maas, H. L. J., and Levine, S. Z. (2019). Extending psychometric network analysis: empirical evidence against g in favor of mutualism? *Intelligence* 73, 52–62. doi: 10.1016/j.intell.2018.12.004

Kane, M. T. (2001). Current concerns in validity theory. *J. Educ. Meas.* 38, 319–342. doi: 10.1111/j.1745-3984.2001.tb01130.x

Kang, Y., Park, J., Yu, K., and Lee, B. C. (2009). A reliability, validity, and normative study of the Korean-Montreal cognitive assessment (K-MoCA) as an instrument for screening of vascular cognitive impairment (VCI). *Korean J. Clin. Psychol.* 28, 549–562. doi: 10.15842/kjcp.2009.28.2.013

Kessels, R. P. C. (2019). Improving precision in neuropsychological assessment: bridging the gap between classic paper-and-pencil tests and paradigms from cognitive neuroscience. *Clin. Neuropsychol.* 33, 357–368. doi: 10.1080/13854046.2018.1518489

Kievit, R. A., Lindenberger, U., Goodyer, I. M., Jones, P. B., Fonagy, P., Bullmore, E. T., et al. (2017). Mutualistic coupling between vocabulary and reasoning supports cognitive development during late adolescence and early adulthood. *Psychol. Sci.* 28, 1419–1431. doi: 10.1177/0956797617710785

Kim, J., Na, H. K., Byun, J., Shin, J., Kim, S., Lee, B. H., et al. (2017). Tracking cognitive decline in amnestic mild cognitive impairment and early-stage Alzheimer dementia: mini-mental state examination versus neuropsychological battery. *Dement. Geriatr. Cogn. Disord.* 44, 105–117. doi: 10.1159/000478520

Kim, Y., Yun, J., Seo, S. W., Kim, J. P., Jang, H., Kim, H. J., et al. (2024). Difference in trajectories according to early amyloid accumulation in cognitively unimpaired elderly. *Eur. J. Neurol.* 31:e16482. doi: 10.1111/ene.16482

Kirchhoff, B. A., Gordon, B. A., and Head, D. (2014). Prefrontal gray matter volume mediates age effects on memory strategies. *NeuroImage* 90, 326–334. doi: 10.1016/J.NEUROIMAGE.2013.12.052

Kiselica, A. M., Webber, T. A., Benge, J. F., Benge, J. F., and Benge, J. F. (2020). The uniform data set 3.0 neuropsychological battery: factor structure, invariance testing, and demographically-adjusted factor score calculation. *J. Int. Neuropsychol. Soc.* 26, 576–586. doi: 10.1017/S135561772000003X

Knight, R. A., and Silverstein, S. M. (2001). A process-oriented approach for averting confounds resulting from general performance deficiencies in schizophrenia. *J. Abnorm. Psychol.* 110, 15–30. doi: 10.1037/0021-843X.110.1.15

Kohoutová, L., Heo, J., Cha, S., Lee, S., Moon, T., Wager, T. D., et al. (2020). Toward a unified framework for interpreting machine-learning models in neuroimaging. *Nat. Protoc.* 15, 1399–1435. doi: 10.1038/s41596-019-0289-5

Kumar, S., Oh, I., Schindler, S., Lai, A. M., Payne, P. R. O., and Gupta, A. (2021). Machine learning for modeling the progression of Alzheimer disease dementia using clinical data: a systematic literature review. *JAMIA Open* 4:ooab052. doi: 10.1093/jamiaopen/ooab052

Kwak, S., Kim, H., Kim, H., Youm, Y., and Chey, J. (2021a). Distributed functional connectivity predicts neuropsychological test performance among older adults. *Hum. Brain Mapp.* 42, 3305–3325. doi: 10.1002/hbm.25436

Kwak, S., Oh, D. J., Jeon, Y. J., Oh, D. Y., Park, S. M., Kim, H., et al. (2022). Utility of machine learning approach with neuropsychological tests in predicting functional impairment of Alzheimer's disease. *J. Alzheimer's Dis.* 85, 1357–1372. doi: 10.3233/JAD-215244

Kwak, S., Park, S. M., Jeon, Y. J., Ko, H., Oh, D. J., and Lee, J. Y. (2021b). Multiple cognitive and behavioral factors link association between brain structure and functional impairment of daily instrumental activities in older adults. *J. Int. Neuropsychol. Soc.* 28, 673–686. doi: 10.1017/S1355617721000916

Kwak, S., You, K., Su, K., Park, M., Young, D., Hairin, O., et al. (2023). Clinical validity of neuropsychological assessment in dementia: a univariate and multivariate methodological comparison. *Korean J. Clin. Psychol.* 42, 118–129. doi: 10.15842/kjcp.2023.42.4.003

Lai, C. (2014). The merits and problems of neuropsychiatric inventory as an assessment tool in people with dementia and other neurological disorders. *Clin. Interv. Aging* 9, 1051–1061. doi: 10.2147/CIA.S63504

Levin, F., Jelistratova, I., Betthauser, T. J., Okonkwo, O., Johnson, S. C., Teipel, S. J., et al. (2021). In vivo staging of regional amyloid progression in healthy middle-aged to older people at risk of Alzheimer's disease. *Alzheimer's Res Ther* 13:178. doi: 10.1186/s13195-021-00918-0

Lilienfeld, S. O., and Strother, A. N. (2020). Psychological measurement and the replication crisis: four sacred cows. *Can. Psychol./Psychol. Can.* 61, 281–288. doi: 10.1037/cap0000236

Lockhart, S. N., and DeCarli, C. (2014). Structural imaging measures of brain aging. *Neuropsychol. Rev.* 24, 271–289. doi: 10.1007/s11065-014-9268-3

Lockhart, S. N., Mayda, A. B. V., Roach, A. E., Fletcher, E., Carmichael, O., Maillard, P., et al. (2012). Episodic memory function is associated with multiple measures of white matter integrity in cognitive aging. *Front. Hum. Neurosci.* 6, 1–12. doi: 10.3389/fnhum.2012.00056

Loewenstein, D. A., Curiel, R. E., Greig, M. T., Bauer, R. M., Rosado, M., Bowers, D., et al. (2016). A novel cognitive stress test for the detection of preclinical Alzheimer disease: discriminative properties and relation to amyloid load. *Am. J. Geriatr. Psychiatry* 24, 804–813. doi: 10.1016/J.JAGP.2016.02.056

Lowry, E., Puthusseryppady, V., Johnen, A.-K., Renoult, L., and Hornberger, M. (2021). Cognitive and neuroimaging markers for preclinical vascular cognitive impairment. *Cereb. Circ. Cogn. Behav.* 2:100029. doi: 10.1016/J.CCCB.2021.100029

Malloy, P. F., and Richardson, E. D. (1994). Assessment of frontal lobe functions. *J. Neuropsychiatry Clin. Neurosci.* 6, 399–410. doi: 10.1176/jnp.6.4.399

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., et al. (2011). 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.* 7, 263–269. doi: 10.1016/j.jalz.2011.03.005

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, a. H., Howerter, A., and Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. *Cogn. Psychol.* 41, 49–100. doi: 10.1006/cogp.1999.0734

Mormino, E. C., Papp, K. V., Rentz, D. M., Donohue, M. C., Amariglio, R., Quiroz, Y. T., et al. (2017). Early and late change on the preclinical Alzheimer's cognitive

composite in clinically normal older individuals with elevated amyloid β. *Alzheimers Dement.* 13, 1004–1012. doi: 10.1016/j.jalz.2017.01.018

Morrison, C., and Oliver, M. D. (2023). Subjective cognitive decline is associated with lower baseline cognition and increased rate of cognitive decline. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 78, 573–584. doi: 10.1093/geronb/gbac178

Mortamais, M., Ash, J. A., Harrison, J., Kaye, J., Kramer, J., Randolph, C., et al. (2017). Detecting cognitive changes in preclinical Alzheimer's disease: a review of its feasibility. *Alzheimers Dement.* 13, 468–492. doi: 10.1016/j.jalz.2016.06.2365

Nation, D. A., Ho, J. K., Dutt, S., Han, S. D., Lai, M. H. C., and Bondi, M. (2019). Neuropsychological decline improves prediction of dementia beyond Alzheimer's disease biomarker and mild cognitive impairment diagnoses. *J. Alzheimer's Dis.* 69, 1171–1182. doi: 10.3233/JAD-180525

Noel-Storr, A. H., McCleery, J. M., Richard, E., Ritchie, C. W., Flicker, L., Cullum, S. J., et al. (2014). Reporting standards for studies of diagnostic test accuracy in dementia: the STARDdem initiative. *Neurology* 83, 364–373. doi: 10.1212/WNL.0000000000000621

Noh, Y., Jeon, S., Lee, J. M., Seo, S. W., Kim, G. H., Cho, H., et al. (2014). Anatomical heterogeneity of Alzheimer disease. *Neurology* 83, 1936–1944. doi: 10.1212/WNL.0000000000001003

Overdorp, E. J., Kessels, R. P. C., Claassen, J. A., and Oosterman, J. M. (2016). The combined effect of neuropsychological and neuropathological deficits on instrumental activities of daily living in older adults: a systematic review. *Neuropsychol. Rev.* 26, 92–106. doi: 10.1007/s11065-015-9312-y

Palliya Guruge, C., Oviatt, S., Delir Haghighi, P., and Pritchard, E. (2021). "Advances in multimodal behavioral analytics for early dementia diagnosis: a review" in Proceedings of the 2021 international conference on multimodal interaction, New York, NY, USA: Association for Computing Machinery (ACM). 328–340.

Papp, K. V., Kaplan, R. F., Springate, B., Moscufo, N., Wakefield, D. B., Guttmann, C. R. G., et al. (2014). Processing speed in normal aging: effects of white matter hyperintensities and hippocampal volume loss. *Neuropsychol. Dev. Cogn. B Aging Neuropsychol. Cogn.* 21, 197–213. doi: 10.1080/13825585.2013.795513

Park, L. Q., Gross, A. L., McLaren, D. G., Pa, J., Johnson, J. K., Mitchell, M., et al. (2012). Confirmatory factor analysis of the ADNI neuropsychological battery. *Brain Imaging Behav.* 6, 528–539. doi: 10.1007/s11682-012-9190-3

Parks, C. M., Iosif, A.-M., Farias, S., Reed, B., Mungas, D., and DeCarli, C. (2011). Executive function mediates effects of white matter hyperintensities on episodic memory. *Neuropsychologia* 49, 2817–2824. doi: 10.1016/j.neuropsychologia.2011.06.003

Pawlowski, J., Segabinazi, J. D., Wagner, F., and Bandeira, D. R. (2013). A systematic review of validity procedures used in neuropsychological batteries. *Psychol. Neurosci.* 6, 311–329. doi: 10.3922/j.PSNS.2013.3.09

Perrotin, A., de Flores, R., Lamberton, F., Poisnel, G., La Joie, R., de la Sayette, V., et al. (2015). Hippocampal subfield volumetry and 3D surface mapping in subjective cognitive decline. *J. Alzheimer's Dis.* 48, S141–S150. doi: 10.3233/JAD-150087

Petersen, R. C., Roberts, R. O., Knopman, D. S., Boeve, B. F., Geda, Y. E., Ivnik, R. J., et al. (2009). Mild cognitive impairment: ten years later. *Arch. Neurol.* 66, 1447–1455. doi: 10.1001/ARCHNEUROL.2009.266

Petersen, R. C., Wiste, H. J., Weigand, S. D., Fields, J. A., Geda, Y. E., Graff-Radford, J., et al. (2021). NIA-AA Alzheimer's disease framework: clinical characterization of stages. *Ann. Neurol.* 89, 1145–1156. doi: 10.1002/ana.26071

Phillips, L. (2004). "Do "frontal tests" measure executive function? Issues of assessment and evidence from fluency tests" in Methodology of frontal and executive function. (London: Taylor and Francis), 185–207.

Pichet Binette, A., Theaud, G., Rheault, F., Roy, M., Collins, D. L., Levin, J., et al. (2021). Bundle-specific associations between white matter microstructure and aβ and tau pathology in preclinical Alzheimer's disease. *eLife* 10:e62929. doi: 10.7554/eLife.62929

Prins, N. D., and Scheltens, P. (2015). White matter hyperintensities, cognitive impairment and dementia: an update. *Nat. Rev. Neurol.* 11, 157–165. doi: 10.1038/nrneurol.2015.10

Reitan, R. M., and Wolfson, D. (2009). "The Halstead-Reitan neuropsychological test battery for adults — theoretical, methodological, and validational bases" in Neuropsychological assessment of neuropsychiatric and neuromedical disorders. 3rd ed. New York, NY, USA: Oxford University Press.

Rotenberg, S., Ruthralingam, M., Hnatiw, B., Neufeld, K., Yuzwa, K. E., Arbel, I., et al. (2020). Measurement properties of the multiple errands test: a systematic review. *Arch. Phys. Med. Rehabil.* 101, 1628–1642. doi: 10.1016/j.apmr.2020.01.019

Royall, D. R., and Palmer, R. F. (2012). Getting past "g": testing a new model of dementing processes in persons without dementia. *J. Neuropsychiatry Clin. Neurosci.* 24, 37–46. doi: 10.1176/appi.neuropsych.11040078

Sachdev, P. S., Blacker, D., Blazer, D. G., Ganguli, M., Jeste, D. V., Paulsen, J. S., et al. (2014). Classifying neurocognitive disorders: the DSM-5 approach. *Nat. Rev. Neurol.* 10, 634–642. doi: 10.1038/nrneurol.2014.181

Sbordone, R. J. (1997). Ecological validity of neuropsychological testing. eds. Sbordone and Long. Delray Beach, FL, USA: GR Press/St. Lucie Press.

Schulz, M. A., Yeo, B. T. T., Vogelstein, J. T., Mourao-Miranda, J., Kather, J. N., Kording, K., et al. (2020). Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nat. Commun.* 11, 1–15. doi: 10.1038/s41467-020-18037-z

Seghier, M. L., and Price, C. J. (2018). Interpreting and Utilising Intersubject variability in brain function. *Trends Cogn. Sci.* 22, 517–530. doi: 10.1016/j.tics.2018.03.003

Shao, Z., Janse, E., Visser, K., and Meyer, A. S. (2014). What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Front. Psychol.* 5:772. doi: 10.3389/fpsyg.2014.00772

Siedlecki, K. L., Stern, Y., Reuben, A., Sacco, R. L., Elkind, M. S. V., and Wright, C. B. (2009). Construct validity of cognitive reserve in a multiethnic cohort: the northern Manhattan study. *J. Int. Neuropsychol. Soc.* 15, 558–569. doi: 10.1017/S1355617709090857

Siegel, J. S., Ramsey, L. E., Snyder, A. Z., Metcalf, N. V., Chacko, R. V., Weinberger, K., et al. (2016). Disruptions of network connectivity predict impairment in multiple behavioral domains after stroke. *Proc. Natl. Acad. Sci.* 113, E4367–E4376. doi: 10.1073/pnas.1521083113

Smith, E. E., Salat, D. H., Jeng, J., McCreary, C. R., Fischl, B., Schmahmann, J. D., et al. (2011). Correlations between MRI white matter lesion location and executive function and episodic memory. *Neurology* 76, 1492–1499. doi: 10.1212/WNL.0b013e318217e7c8

Soldan, A., Pettigrew, C., Fagan, A. M., Schindler, S. E., Moghekar, A., Fowler, C., et al. (2019). ATN profiles among cognitively normal individuals and longitudinal cognitive outcomes. *Neurology* 92, e1567–e1579. doi: 10.1212/WNL.0000000000007248

Sperber, C., Griffis, J., and Kasties, V. (2022). Indirect structural disconnection-symptom mapping. *Brain Struct. Funct.* 227, 3129–3144. doi: 10.1007/s00429-022-02559-x

Sperber, C., and Karnath, H. O. (2018). On the validity of lesion-behaviour mapping methods. *Neuropsychologia* 115, 17–24. doi: 10.1016/J.NEUROPSYCHOLOGIA.2017.07.035

Stern, Y., Habeck, C., Steffener, J., Barulli, D., Gazes, Y., Razlighi, Q., et al. (2014). The reference ability neural network study: motivation, design, and initial feasibility analyses. *NeuroImage* 103, 139–151. doi: 10.1016/j.neuroimage.2014.09.029

Strauss, M. E., and Smith, G. T. (2009). Construct validity: advances in theory and methodology. *Annu. Rev. Clin. Psychol.* 5, 1–25. doi: 10.1146/annurev.clinpsy.032408.153639

Tetreault, A. M., Phan, T., Orlando, D., Lyu, I., Kang, H., Landman, B., et al. (2020). Network localization of clinical, cognitive, and neuropsychiatric symptoms in Alzheimer's disease. *Brain* 143, 1249–1260. doi: 10.1093/BRAIN/AWAA058

Teuber, H. L. (1955). Physiological psychology. *Annu. Rev. Psychol.* 6, 267–296. doi: 10.1146/annurev.ps.06.020155.001411

Thomas, K. R., Bangen, K. J., Weigand, A. J., Edmonds, E. C., Wong, C. G., Cooper, S., et al. (2020). Objective subtle cognitive difficulties predict future amyloid accumulation and neurodegeneration. *Neurology* 94, e397–e406. doi: 10.1212/WNL.0000000000008838

Thomas, K. R., Eppig, J. S., Weigand, A. J., Edmonds, E. C., Wong, C. G., Jak, A. J., et al. (2019). Artificially low mild cognitive impairment to normal reversion rate in the Alzheimer's Disease Neuroimaging Initiative. *Alzheimers Dement.* 15, 561–569. doi: 10.1016/j.jalz.2018.10.008

Treviño, M., Zhu, X., Lu, Y. Y., Scheuer, L. S., Passell, E., Huang, G. C., et al. (2021). How do we measure attention? Using factor analysis to establish construct validity of neuropsychological tests. *Cogn. Res. Princ. Implic.* 6, 1–26. doi: 10.1186/s41235-021-00313-1

Tsoi, K. F., Chan, J. Y. C., Hirai, H. W., Wong, S. Y. S., and Kwok, T. C. Y. (2015). Cognitive tests to detect dementia. *JAMA Intern. Med.* 175:1450. doi: 10.1001/jamainternmed.2015.2152

Tucker-Drob, E. M. (2019). Cognitive aging and dementia: a life-span perspective. *Ann. Rev. Dev. Psychol.* 1, 177–196. doi: 10.1146/annurev-devpsych-121318-085204

Vaidya, A. R., Pujara, M. S., Petrides, M., Murray, E. A., and Fellows, L. K. (2019). Lesion studies in contemporary neuroscience. *Trends Cogn. Sci.* 23, 653–671. doi: 10.1016/j.tics.2019.05.009

Van Der Maas, H., Kan, K.-J., Marsman, M., and Stevenson, C. E. (2017). Network models for cognitive development and intelligence. *J. Intelligence* 5:16. doi: 10.3390/jintelligence5020016

Vasquez, B. P., and Zakzanis, K. K. (2015). The neuropsychological profile of vascular cognitive impairment not demented: a meta-analysis. *J. Neuropsychol.* 9, 109–136. doi: 10.1111/jnp.12039

Vogel, J. W., Young, A. L., Oxtoby, N. P., Smith, R., Ossenkoppele, R., Strandberg, O. T., et al. (2021). Four distinct trajectories of tau deposition identified in Alzheimer's disease. *Nat. Med.* 27, 871–881. doi: 10.1038/s41591-021-01309-6

Westen, D., and Rosenthal, R. (2003). Quantifying construct validity: two simple measures. *J. Pers. Soc. Psychol.* 84, 608–618. doi: 10.1037/0022-3514.84.3.608

Weuve, J., Proust-Lima, C., Power, M. C., Gross, A. L., Hofer, S. M., Thiébaut, R., et al. (2015). Guidelines for reporting methodological challenges and evaluating

potential bias in dementia research for the MELODEM initiative 3 HHS public access. *Alzheimers Dement.* 11, 1098–1109. doi: 10.1016/j.jalz.2015.06.1885

Whiteside, D. M., Kealey, T., Semla, M., Luu, H., Rice, L., Basso, M. R., et al. (2016). Verbal fluency: language or executive function measure? *Appl. Neuropsychol. Adult* 23, 29–34. doi: 10.1080/23279095.2015.1004574

Whitwell, J. L., Przybelski, S. A., Weigand, S. D., Ivnik, R. J., Vemuri, P., Gunter, J. L., et al. (2009). Distinct anatomical subtypes of the behavioural variant of frontotemporal dementia: a cluster analysis study. *Brain* 132, 2932–2946. doi: 10.1093/brain/awp232

Wolfsgruber, S., Jessen, F., Wiese, B., Stein, J., Bickel, H., Mösch, E., et al. (2014). The CERAD neuropsychological assessment battery total score detects and predicts alzheimer disease dementia with high diagnostic accuracy. *Am. J. Geriatr. Psychiatry* 22, 1017–1028. doi: 10.1016/j.jagp.2012.08.021

Yarkoni, T., and Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspect. Psychol. Sci.* 12, 1100–1122. doi: 10.1177/1745691617693393

Yonelinas, A. P., Ramey, M. M., Riddell, C., Kahana, M. J., and Wagner, A. D. (2022). Recognition memory: the role of recollection and familiarity. New York, NY: Oxford University Press.

Yonelinas, A. P., Widaman, K., Mungas, D., Reed, B., Weiner, M. W., and Chui, H. C. (2007). Memory in the aging brain: doubly dissociating the contribution of the hippocampus and entorhinal cortex. *Hippocampus* 17, 1134–1140. doi: 10.1002/hipo.20341

Younes, L., Albert, M., Moghekar, A., Soldan, A., Pettigrew, C., and Miller, M. I. (2019). Identifying changepoints in biomarkers during the preclinical phase of Alzheimer's disease. *Front. Aging Neurosci.* 11:74. doi: 10.3389/fnagi.2019.00074