



Research Article

Psychometric modeling of Boston process approach data for dementia prediction in the Framingham Heart Study

Brandon Frank^{1,2,3} , Ashita Gurnani^{3,4}, Landon Hurley¹, Calvin Guan⁴, Stacy L. Andersen⁵ , Sherral A. Devine^{4,6}, Maureen K. O'Connor^{2,3,7}, Andrew Budson^{1,2,3}, Chunyu Liu⁸, Honghuang Lin⁹, Sanford Auerbach^{3,4}, Yulin Liu^{3,4}, David J Libon¹⁰ , Catherine C. Price¹¹, Lindsay Farrer^{2,4}, Jesse Mez^{2,3,4}, Alvin Ang^{4,6,12} and Rhoda Au^{2,3,4,6,12,13}

¹Neurology, VA Boston Healthcare System, USA, ²Alzheimer's Disease Research Center, Boston University Chobanian & Avedisian School of Medicine, USA, ³Department of Neurology, Boston University Chobanian & Avedisian School of Medicine, USA, ⁴The Framingham Heart Study, USA, ⁵Department of Medicine, Boston University Chobanian & Avedisian School of Medicine, USA, ⁶Department of Anatomy and Neurobiology, Boston University Chobanian & Avedisian School of Medicine, USA, ⁷Psychology, VA Bedford Healthcare System, USA, ⁸Department of Biostatistics, Boston University School of Public Health, ⁹Department of Medicine, University of Massachusetts Chan Medical School, USA, ¹⁰Geriatric and Gerontology, New Jersey Institute for Successful Aging, and Department of Psychology, Rowan University School of Osteopathic Medicine, USA, ¹¹Clinical and Health Psychology, University of Florida, USA, ¹²Slope Epidemiology Center, Boston University Chobanian & Avedisian School of Medicine, USA and ¹³Department of Epidemiology, Boston University School of Public Health, USA

Abstract

Objectives: Neuropsychological (NP) tests are multi-domain in execution. Reliance on a single score representing specific domains obscures the detection of subtle cognitive changes and increases risk of inaccurate assessment. Rooted in the Boston Process Approach (BPA), the Framingham Heart Study (FHS) captures multi-dimensional errors and process features within and across NP tests. We examined these BPA variables in community-dwelling older adults. **Methods:** We analyzed data from 2363 dementia-free participants aged 60 and above. Exploratory and confirmatory factor analyses used Kemeny covariance structures. Measurement invariance was estimated across age, sex, and education groups. We assessed the impact of demographics on latent factors, and the ability of these factors to predict future conversion to all-cause dementia. We trained machine learning (ML) models to compare NP and BPA data. **Results:** Participants were older adults (mean age 71.5 ± 8.7 years), primarily female (54.2%), and non-Hispanic White (96.5%). The bifactor model was the only model with adequate fit (CFI = 0.96, RMSEA = 0.03). General and specific factors captured ability for accurate and strategic responses, test-specific variance, and nuanced executive and semantic processes distributed across tests. Higher general ability and stronger verbatim story recall were associated with a reduced likelihood of developing all-cause dementia (general: OR = 0.15, 95% CI [0.12–0.86]; recall: OR = 0.24, 95% CI [0.23–0.90]) over a median of 5.2 years. With NP/BPA data, ML models identified >99% of 222 converters. **Conclusions:** This study highlights the strengths of NP/BPA data. Multidimensional cognitive features may enhance sensitivity to early changes predictive of incipient dementia.

Keywords: Boston process approach; dementia; Framingham Heart Study; machine learning; measurement invariance; neuropsychology; structural equation modeling

(Received 11 July 2025; final revision 16 February 2026; accepted 24 February 2026; First Published online 14 April 2026)

Statement of Research Significance

Research Question(s) or Topic(s): We examined patterns of errors and response strategies used by older adults from the Framingham Heart Study during cognitive tasks. We investigated whether these behaviors reflect underlying cognitive abilities, and the relationships between these abilities and demographics. We also examined whether these abilities identify persons that will develop dementia. **Main Findings:** We found that errors and response strategies reflected a general ability, but also seven underlying cognitive processes that were related to each of the cognitive tasks. This pattern was consistent across important demographics like age, sex, and education. Older, less educated participants produced fewer accurate and strategic responses, while differences between males and females were varied. Importantly, this approach was helpful in identifying participants that would develop dementia. **Study**
(Continued)

Corresponding author: Brandon Frank; Email: befrank1@bu.edu

Brandon Frank and Ashita Gurnani Indicates shared first authorship. These authors contributed equally to the manuscript.

Cite this article: Frank B., Gurnani A., Hurley L., Guan C., Andersen S.L., Devine S.A., O'Connor M.K., Budson A., Liu C., Lin H., Auerbach S., Liu Y., Libon D.J., Price C.C., Farrer L., Mez J., Ang A., & Au R. (2026) Psychometric modeling of Boston process approach data for dementia prediction in the Framingham Heart Study. *Journal of the International Neuropsychological Society*, 32: 243–254, <https://doi.org/10.1017/S1355617726101921>

© The Author(s), 2026. Published by Cambridge University Press on behalf of International Neuropsychological Society. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Statement of Research Significance (Continued)

Contributions: This study provides an expanded approach for researchers to use cognitive tests in studies of aging but also reinforces how clinicians typically understand test performance.

Introduction

Summary scores from neuropsychological (NP) assessments are widely used to infer underlying cognitive abilities (Au et al., 2025; Casaletto & Heaton, 2017; Libon et al., 2025; Salmon & Bondi, 2009). However, a single score representing a specific cognitive domain often obscures the diverse cognitive strategies individuals may employ to complete any given NP test. Reliance on summary scores can obscure subtle cognitive changes, which is especially critical for early detection of cognitive disorders (Kaplan, 1988; Libon et al., 2022, 2025). Furthermore, although NP tests are inherently multi-domain, many studies artificially simplify their structure by assigning each test to a single cognitive domain (Au & Devine, 2013). This “one test-one domain” approach can result in misinterpretation of normal performance variability as cognitive impairment (Au et al., 2025; Jak et al., 2009; Loewenstein et al., 2007).

The Boston Process Approach (BPA), created by Edith Kaplan (Kaplan, 1988, 1990; Libon et al., 2022), provides a solution to the “one test-one domain” problem by analyzing and quantifying the types of cognitive resources utilized and errors made during the process of completing a neurocognitive operation. Similarly, while a summary score may indicate presence of impairment, using the BPA further elucidates specific brain-behavior relationships to better inform level of impairment and differential diagnosis, including early detection of subtle cognitive changes. Research using BPA has largely focused on differentiating between clinical groups (Bélanger & Belleville, 2009; Bettcher et al., 2008; Libon et al., 2018) as well as early detection of MCI (Eppig et al., 2012; Libon et al., 2011). For example, early studies used the BPA to distinguish Alzheimer’s disease (AD) and vascular dementia (VaD), with the former group largely producing errors associated with amnesic and lexical impairments, whereas the latter group producing executive dysfunction errors across many cognitive domains (Lamar et al., 1997, 2010; Libon et al., 2018).

The emerging landscape of approved anti-amyloid drugs for treatment of Alzheimer’s disease and related dementias (ADRD) has pushed for the development of new methodologies that can detect subtle cognitive changes associated with MCI and dementia when disease modifying therapies are most effective (Au et al., 2025; Cummings et al., 2024; Libon et al., 2022, 2025; Libon, Swenson, Price, et al., 2024). There is emerging evidence that digital NP assessment, which is also rooted in BPA, validly captures drawn, written, and spoken output from NP tests in real time to identify markers of subtle cognitive impairment (Libon et al., 2022, 2025; Libon, Swenson, Price, et al., 2024; Libon, Swenson, Tobyne, et al., 2024). This includes research using the digital clock drawing test (dCDT) (Binaco et al., 2020; Davoudi et al., 2021; Dion et al., 2021; Frank et al., 2024; Müller et al., 2019; Zheng et al., 2021) and artificial intelligence (AI)-assisted models extracting linguistic and acoustic features from speech produced during NP testing (Amini et al., 2024; Ding et al., 2024; Tavabi et al., 2022; Young et al., 2024).

While digital NP assessment holds tremendous potential, this line of research is in nascent stages and its clinical utility and scalability, at least currently, is unknown (Au et al., 2023; Libon et al., 2022, 2025). On the other hand, BPA metrics derived from traditional NP tests are routinely utilized in clinical evaluation and

relatively better understood, however, they are considered non-standardized and case-specific, subjective, or difficult to systematically quantify without expert review (Au & Devine, 2013). As such, use of BPA in research has been limited by adoption constraints, including time, labor, and intensive training.

The current study focuses on NP data from the Framingham Heart Study (FHS), a community-based cohort study established to examine cardiovascular disease, which has collected comprehensive BPA data since 2005 using an extensive scoring system and weekly meetings to maintain reliability. These data include distinct process variables related to errors, approaches, strategies, and “testing the limits” procedures. Several studies have examined FHS BPA data, especially using the clock drawing test, and demonstrated its clinical and diagnostic utility (Gupta et al., 2015; Hanke et al., 2013; Libon et al., 2015; Nishtala et al., 2014).

Despite the richness of BPA data, there remain methodological challenges. Most variables are zero-inflated counts, with each participant having a distinct pattern of errors and approaches. Any particular response type or error has a relatively low frequency but is often clinically meaningful. As a result, these data are inappropriate for traditional psychometric methods and analyses. Ferretti et al. (2024) recently examined the ability of 23 BPA variables to predict conversion to dementia in an FHS sample, stratified by sex. They found associations between derived BPA composites and conversion, which varied by sex. However, there is a need to expand on this study with additional psychometric investigation (e.g., to empirically assess constituent BPA data, and to assess factor structure and measurement invariance (MI); see Measurement and Latent Factor section below).

Overall, given the richness and complexity of the BPA data collected at FHS, psychometric analyses are needed to (1) understand what these variables measure in isolation or aggregation for this sample of older adults and (2) aggregate the data and identify underlying latent cognitive structures. In this study, we conducted psychometric analyses of constituent BPA variables (i.e., individual variables that were derived from detailed scoring methods but not aggregated with any theoretically driven process). We employed a novel method appropriate for the sparsity and complexity of the data. This included examination of factor structure, MI between demographic groups, and the utility of our retained bifactor model to predict dementia. Using the best-fitting models, we examined the influence of demographic factors on latent cognitive components and assessed their predictive utility for future conversion to dementia. We contend this study will inform clinical understanding of long-used NP tests and display a methodological approach for future studies. This could serve as a foundation to further demonstrate the utility of BPA data in predicting outcomes related to aging and dementia.

Methods

Participants

The FHS is a multi-generational longitudinal community study including Gen 1 (initiated in 1948), Gen 2 (1971), Gen 3 (2002), Omni 1 (1994), and Omni 2 (2003) cohorts. NP testing was first

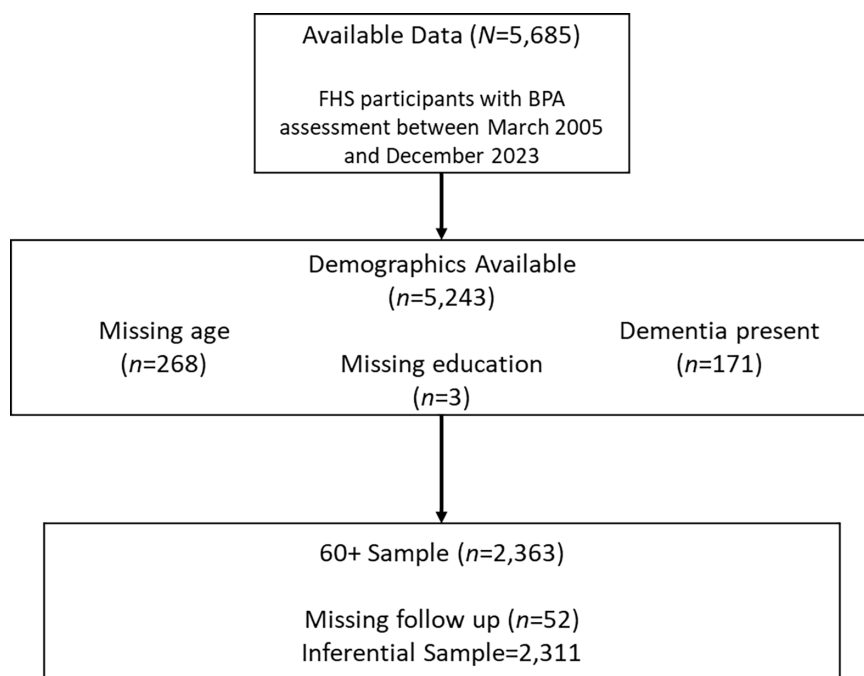


Figure 1. CONSORT flow diagram for sample selection process. Boston process approach (BPA); Framingham Heart Study (FHS).

implemented in 1989; BPA scoring was initiated in 2005, resulting in a cross-sectional sample of 5685 participants with available BPA data (Au & Devine, 2013; Ferretti et al., 2024). Given our primary interest in aging and dementia, the analytical sample focused on adults ≥ 60 years ($n = 2363$). However, a Supplement presents a discussion and analyses of all participants with available data unrestricted by age. We also examined an alternate age cutoff for older adults with MI analyses (see Analytic Plan). Participants were excluded if there was missing information on target demographic variables, including age at BPA evaluation ($n = 268$) and education ($n = 3$). Following Ferretti et al. (2024), 171 participants with diagnosed dementia at the first available BPA assessment were excluded, as these individuals tend to discontinue from tasks, resulting in a high degree of missing data. To maintain consistency with this prior study, we included 108 participants presenting for BPA assessment with non-dementia cognitive impairment (i.e., mild cognitive impairment; MCI). We report on these participants in the dementia prediction section below. Of note, even MCI patients with an underlying neurodegenerative disease would still be considered in early stages of the condition, and therefore relevant for efforts at early detection. In dementia prediction models, participants were excluded if they did not receive at least one longitudinal follow-up including NP evaluation after their initial baseline assessment ($n = 52$). This led to cohorts of 2363 for psychometric models and 2311 for the inferential model (see analytic plan). See Figure 1 for the Consort Flow diagram of the sample selection process. Involvement of human subjects was done in accord with the Helsinki Declaration of 1975. This study was approved by the institutional review board at Boston University. All participants provided written informed consent.

Neuropsychological testing

The FHS NP battery has been described in detail (Au et al., 2004). FHS participants are invited to complete comprehensive NP testing at least every 4–5 years. The battery includes the following tests: Wechsler Memory Scale-First Edition (WMS): Logical

Memory (LM)–Passage A, Visual Reproduction (VR), and Paired Associate Learning (PAL) (Wechsler, 1945), Wechsler Adult Intelligence Scale First Edition (WAIS): Digit Span and Similarities (Wechsler, 1955), WAIS-IV Coding (Wechsler, 2008); Boston Naming Test (BNT) (Kaplan et al., 1983), Trail-Making Test Parts A and B (Trails A and B), Finger Tapping, Verbal Fluency (Spree & Strauss, 1991), Clock Drawing Test (Kaplan, 1988, 1990), and Hooper Visual Organization Test (Hooper, 1983). Of note, some of these tests use idiosyncratic instructions, such as the instruction to not lift the pen from the page on Trails A and B. Using the BPA, errors and extraneous responses were quantified from each test including but not limited to confabulations, interference, contaminations, and perseverations. The FHS battery and BPA scoring criteria permit quantification of 169 distinct variables from the various tests. Of these variables, only 35 variables were retained in the primary model (see factor analyses below). We describe each of the variables in the Supplement (see Supplemental Table 2a–c).

Diagnosis

Surveillance of incident dementia began in 1975. Dementia diagnoses were determined through FHS consensus adjudication, involving review by at least one neurologist and one neuropsychologist, based on neurological and NP assessments, medical records, FHS health examinations, and in some cases, family interviews. Details of the surveillance and diagnosis of dementia within FHS have been published previously (Kolo et al., 2025; Satizabal et al., 2016). FHS diagnostic criteria for dementia follow the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (DSM-IV).

Measurements and latent factors

Because of the high number of available BPA variables ($n = 169$), there is a need for data reduction to permit interpretable findings. However, simply combining counts of errors from within or

between tests would not necessarily satisfy model assumptions because the latent structure and model fit would be unexamined (Borsboom, 2006). While it is typical to aggregate BPA variables based on theory and/or past research, there remains a need for empirical studies of unaggregated measures. Factor analysis provides a means to explore empirical associations between items and measures. These findings can be used in structural models for inference. We assume that all performances on cognitive tests, and by extension BPA data, reflect a combination of latent cognitive abilities/processes and noise/error. Traditional analyses do not permit examination or differentiation of signal and error, which can bias results from models involving cognitive variables (Nichols et al., 2024).

While factor models provide an alternative, this methodology still faces the problem of factor score indeterminacy (Guttman, 1955; McDonald & Bolt, 1998; Rigdon et al., 2019; Waller, 2023), in which any factor solution could produce a multitude of factor scores, resulting in a source of error when these scores are used to create composites outside of the factor structure from which they were derived. The current study employed structural equation models (SEMs), which, while complex, provide an optimal means to examine the predictive value of latent variables (McNeish, 2023). In addition, MI analyses were incorporated to assess whether a particular scale or set of items measured a particular construct in the same way for different groups.

We empirically assessed variable inclusion, factor retention, and factor structure, as well as model fit. For factor solutions that demonstrated superior fit, we assessed relative improvement with a bifactor model (Holzinger & Swineford, 1937). The bifactor model is a traditional means to conceptualize cognitive test performance with a general factor influencing all performances and specific factors reflecting test-specific variance (i.e., unrelated to the general factor). Given the nature of BPA data, including zero-inflation and count distribution, we employed a novel application of Kemeny metric space with maximum likelihood decomposition, which constructs a non-parametric covariance matrix built upon a generalized linear rank framework. This approach is optimal for our data, given affine-linear invariance over monotone transformations, resulting in stable linear decomposition and analysis (Emond & Mason, 2002).

We assessed MI in a SEM framework (Borsboom, 2006; Millsap, 1997) using the traditional four hierarchical levels (i.e., constraining various parameters to be equal and then comparing free and constrained models). First, we assessed the equivalence of the structural model between groups (configural invariance), then the equivalence of factor loadings (weak or metric invariance), then item intercepts (strong or scalar invariance), and finally the residual or unexplained variance (strict invariance).

Analytic plan

Analyses were conducted with the “psyc” and “lavaan” packages in R (Revelle, 2024; Rosseel et al., 2014). All BPA variables were missing at less than 12%, all but three were missing at less than 10%, and 74.8% were missing less than 5% ($\bar{x}_{\text{missing}} = 3.4\%$; median = 3%). Missing values were imputed using multiple imputation via random forests ($k = 5$). Visual inspection of density plots indicated no substantial difference between the distributions of the original and imputed data. To determine the appropriate number of latent factors, we employed several factor retention methods, including parallel analysis (Horn, 1965; Lim & Jahng, 2019), the Minimum Average Partial (MAP) criterion

(Velicer, 1976), and Very Simple Structure (VSS; (Revelle & Rocklin, 1979). Each of these methods has limitations; for example, parallel analysis tends to provide overly complex solutions with larger samples as simulated factors have eigenvalues close to 1 (Tran & Formann, 2009). Since these methods produced divergent results in our study, we estimated the interpretability of various exploratory factor solutions. Model fit was then assessed with confirmatory factor analysis (CFA) in the “lavaan” package. The cut-point for item loadings to be included in confirmatory models was set at 0.3 (Fabrigar et al., 1999). An oblimin rotation was used for all exploratory models because these solutions improved confirmatory fit. Model fit was primarily assessed with the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and the standardized root-mean-square residual (SRMR) (Bentler, 1990; Browne & Cudeck, 1993; Hu & Bentler, 1999), using the following standard thresholds: CFI > 0.95, RMSEA < 0.06, and SRMR < 0.08 (Hu & Bentler, 1999). For MI analyses, groups were constructed based on cutoffs for age at BPA assessment (cutoff at age 65; $n_{<65} = 694$, $n_{\geq 65} = 1669$), sex (male or female; $n_{\text{male}} = 1,082$, $n_{\text{female}} = 1281$), and education (cutoff of a high school diploma; $n_{\leq 12\text{yrs}} = 731$, $n_{> 12\text{yrs}} = 1632$). We did not assess MI with ethnicity because most participants identified as non-Hispanic White. Of note, age groups were chosen to assess the validity of a more liberal cutoff for older adults (i.e., 60+), versus the more traditional cutoff of 65, which allowed us to maximize the sample included. Following (Putnick & Bornstein, 2016) and (Chen, 2007), we assessed MI with χ^2 difference tests and change in fit indices including decline ≥ 0.010 in CFI (ΔCFI), along with increase ≥ 0.015 in RMSEA (ΔRMSEA) or increase ≥ 0.030 in SRMR (ΔSRMR). Given that the model demonstrated adequate MI for age, sex, and education, we assessed the extent to which these demographic factors predicted latent components of the model. In the subsample with follow-up after BPA assessment ($n = 2311$), we also assessed the extent to which latent components predicted future conversion to dementia. We assessed the predictive value of general and specific factors separately given the complexity of the models and interpretation of these factors. However, all other components of the SEMs were held constant, including exogenous predictors (age, sex, and education). Dementia conversion was assessed with probit regression, which allows only approximate conversion to an odds ratio (OR) (Amemiya, 1981). Unless otherwise specified, standardized regression weights are reported. As a final exploratory analysis, we trained machine-learning models with a random forest classifier to predict future conversion to dementia. See Supplemental Methods for details on training, tuning, and model interpretation. Model performance was estimated using five-fold cross-validation and pooled across imputed datasets, yielding accuracy, area under the receiver operating characteristic curve (AUC), area under the precision–recall curve (PR AUC), sensitivity, and specificity (Breiman, 2001). Because dementia conversion was rare (9.4%), PR AUC was selected as the primary performance metric (Saito & Rehmsmeier, 2015). The initial model was trained on traditional NP summary scores. A second model included only BPA variables. Finally, we evaluated a third model with combined NP/BPA data to examine incremental benefit with expanded predictors. To quantify the contribution of each predictor in the combined NP/BPA models, we computed permutation importance values, defined as the reduction in model performance after randomly permuting each variable (Fisher et al., 2019).

Table 1. Demographic information by cohort at Boston process approach assessment

	Gen 1 (<i>n</i> = 157)	Gen 2 (<i>n</i> = 1904)	Gen 3 (<i>n</i> = 254)	Omni 2 (<i>n</i> = 48)
Age (median [IQR])	89.39 [87.76, 91.35]	69.86 [64.72, 76.45]	62.85 [61.23, 65.83]	66.78 [63.46, 74.00]
Sex (% female)	105 (66.9)	1019 (53.5)	132 (52.0)	25 (52.1)
Race/ethnicity (%)				
Non-Hispanic White	157 (100.0)	1875 (98.5)	248 (97.6)	0 (0.0)
Black/African American	0 (0.0)	5 (0.3)	0 (0.0)	14 (29.2)
Asian/Indian/Pacific Islander	0 (0.0)	1 (0.1)	0 (0.0)	21 (43.8)
Hispanic/Latino	0 (0.0)	6 (0.3)	0 (0.0)	9 (18.8)
Others	0 (0.0)	17 (0.9)	6 (2.4)	4 (8.3)
Education (%)				
<High school	29 (18.5)	75 (3.9)	2 (0.8)	2 (4.2)
High school	73 (46.5)	508 (26.7)	40 (15.7)	2 (4.2)
Some college	25 (15.9)	596 (31.3)	83 (32.7)	12 (25.0)
College	30 (19.1)	725 (38.1)	129 (50.8)	32 (66.7)

Note: Includes age at assessment, sex (male or female), race/ethnicity, and educational attainment.

Table 2. Exploratory factor structure statistics (*n* = 2363)

Factor	<i>df</i>	χ^2	<i>p</i>	RMSEA	SRMR	BIC	VSS	MAP
1	14,027.00	13,295.56	1.00	0.00	0.02	−95,661.79	0.15	0.0006
2	13,859.00	9,555.05	1.00	0.00	0.02	−98,097.33	0.17	0.0004
3	13,692.00	8,114.92	1.00	0.00	0.01	−98,240.25	0.16	0.0004
4	13,526.00	7,127.84	1.00	0.00	0.01	−97,937.90	0.16	0.0004
5	13,361.00	6,178.80	1.00	0.00	0.01	−97,605.27	0.16	0.0004
6	13,197.00	5,509.63	1.00	0.00	0.01	−97,000.53	0.17	0.0005
7	13,034.00	4,878.36	1.00	0.00	0.01	−96,365.68	0.17	0.0005
8	12,872.00	4,400.65	1.00	0.00	0.01	−95,585.02	0.18	0.0005
9	12,711.00	3,861.61	1.00	0.00	0.01	−94,873.46	0.16	0.0006
10	12,551.00	3,473.57	1.00	0.00	0.01	−94,018.67	0.13	0.0006
11	12,392.00	3,118.72	1.00	0.00	0.01	−93,138.47	0.16	0.0006
12	12,234.00	2,769.16	1.00	0.00	0.01	−92,260.72	0.17	0.0007
13	12,077.00	2,462.92	1.00	0.00	0.01	−91,347.43	0.17	0.0007
14	11,921.00	2,226.66	1.00	0.00	0.01	−90,371.94	0.18	0.0008
15	11,766.00	2,037.50	1.00	0.00	0.01	−89,357.11	0.18	0.0008

Note: Table displays number of factors retained, degrees of freedom (*df*), chi square (χ^2), probability (*p*), root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR), Bayes Information Criterion (BIC), Very Simple Structure (VSS) fit, and Velicer's Minimum Average Partial (MAP) values.

Results

Demographic characteristics of the sample at baseline BPA assessment are presented in Table 1. The mean age was 71.5 ± 8.7 years old, and 54.2% of participants were female. The majority (95.5%, *n* = 2,255) had completed at least a high school education, and 38.8% held a college degree or higher (*n* = 916). Most participants identified as non-Hispanic White (96.5%, *n* = 2,280). Following the baseline BPA assessment, 9.6% of participants (*n* = 222) were diagnosed with all-cause dementia. Among those who converted, the median time to diagnosis was 5.22 years (interquartile range [IQR]: 2.93–8.54 years). Of the 108 participants who presented to BPA assessment with MCI, fewer than half later converted (46.3%; *n* = 50). Presence of MCI corresponded with a higher rate of conversion but still had poor sensitivity as a marker to correctly identify incipient dementia (only 22.5% of cases).

Factor analyses

The Kaiser-Meyer-Olkin statistic evidenced sampling adequacy, *KMO* = 0.820 (Kaiser, 1974). Bartlett's test of sphericity (Bartlett, 1951) indicated significant correlation between variables, $\chi^2(14196) = 20659.08$, *p* < 0.001. Parallel analysis resulted in retention of

7 factors. The MAP and VSS methods generated retention of 3 factors and 2 factors, respectively. Model fit improved with multidimensionality but without the same linearity. The best-fitting CFA models retained 2 and 7 factors. Of note, higher dimensional models did not converge and fit worsened (i.e., 8 factors and beyond). Bifactor solutions improved model fit for the best-fitting models. These two models retained 24 and 35 variables out of the 169 available BPA variables. The more complex model was chosen to capture more of the available BPA data, and, most importantly, because this was the only model with acceptable fit per a priori decided thresholds, $\chi^2(504) = 928.68$, *p* < 0.001, CFI = 0.97, RMSEA = 0.02, SRMR = 0.02 (see Tables 2 and 3 and Figure 2).

Measurement invariance

Overall, configural, metric, and scalar invariance held across age, sex, and education, but strict invariance did not fully hold for residuals. Regarding age, a multigroup CFA demonstrated configural invariance between age groups (respectively, *n*_{<65} = 694, *n*_{≥65} = 1669; see Table 4). Constraining factor loadings demonstrated metric invariance, $\Delta\chi^2(62) = 27.7$, *p* > 0.999, Δ CFI = 0.001, Δ RMSEA = 0.000, Δ SRMR = 0.000. Constraining item intercepts demonstrated scalar invariance, $\Delta\chi^2(27) = 37.4$, *p* = 0.088,

Table 3. Confirmatory fit statistics ($n = 2,363$)

Model	χ^2	df	p	CFI	TLI	RMSEA	SRMR	BIC
F-1	1,844.05	119	<.001	0.73	0.69	0.08	0.06	386,082.51
F-2	1,362.39	539	<.001	0.94	0.93	0.02	0.02	774,771.93
BIF-2	861.85	227	<.001	0.94	0.92	0.03	0.03	529,948.57
F-3	1,038.61	272	<.001	0.92	0.92	0.04	0.04	553,002.32
F-4	1,485.87	371	<.001	0.91	0.90	0.04	0.04	642,543.55
F-5	1,294.87	341	<.001	0.92	0.91	0.03	0.03	619,650.78
F-6	1,329.80	449	<.001	0.93	0.92	0.03	0.02	709,312.01
F-7	1,362.39	539	<.001	0.94	0.93	0.02	0.02	774,771.93
BIF-7	928.68	504	<.001	0.97	0.96	0.02	0.02	774,610.09
F-8	1,624.47	602	<.001	0.92	0.92	0.03	0.03	820,528.50

Note: Table displays number of factors retained and model type (Factor[F] and Bifactor [BIF], respectively), chi square (χ^2), degrees of freedom (df), probability (p), comparative fit index (CFI), Tucker–Lewis index (TLI), root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR), and Bayes Information Criterion (BIC). Bold values indicate the model with best fit per a priori thresholds.

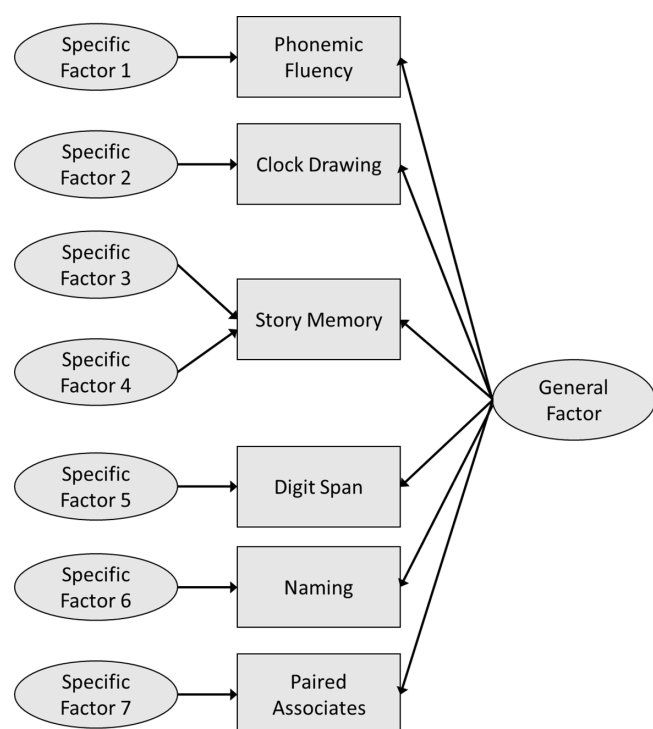


Figure 2. Schematic Display of the Bifactor Model in the 60+ Sample ($n = 2,363$). Note: Individual items and loadings are omitted for visualization. See Supplemental Table 2a for a description of all items, Supplemental Table 5b for loadings, and Supplemental Table 6 for covariance between factors.

$\Delta CFI = 0.000$, $\Delta RMSEA = 0.000$, $\Delta SRMR = 0.002$. However, constraining residuals did not demonstrate strict invariance between age groups, $\Delta\chi^2(35) = 4605.9$, $p < 0.001$, $\Delta CFI = 0.373$, $\Delta RMSEA = 0.058$, $\Delta SRMR = 0.047$. Regarding sex (male or female), a multigroup CFA demonstrated configural invariance (respectively, $n_{\text{male}} = 1,082$, $n_{\text{female}} = 1281$; see Table 4). Constraining factor loadings demonstrated metric invariance, $\Delta\chi^2(62) = 26.6$, $p = 1.000$, $\Delta CFI = 0.001$, $\Delta RMSEA = 0.000$, $\Delta SRMR = 0.001$, and constraining item intercepts demonstrated scalar invariance, $\Delta\chi^2(27) = 13.8$, $p = 0.983$, $\Delta CFI = 0.000$, $\Delta RMSEA = 0.000$, $\Delta SRMR = 0.000$. However, constraining residuals demonstrated questionable strict invariance between males and females, $\Delta\chi^2(35) = 252.9$, $p < 0.001$, $\Delta CFI = 0.015$, $\Delta RMSEA = 0.012$, $\Delta SRMR = 0.004$. Finally, regarding education (high school diploma

or less versus some college or a college degree), a multigroup CFA demonstrated configural invariance (respectively, $n_{\leq 12\text{yrs}} = 731$, $n_{> 12\text{yrs}} = 1632$; see Table 4). Constraining factor loadings demonstrated metric invariance, $\Delta\chi^2(62) = 30.6$, $p = 1.000$, $\Delta CFI = 0.001$, $\Delta RMSEA = 0.000$, $\Delta SRMR = 0.001$, and constraining item intercepts demonstrated adequate scalar invariance, $\Delta\chi^2(27) = 42.1$, $p = 0.032$, $\Delta CFI = 0.000$, $\Delta RMSEA = 0.001$, $\Delta SRMR = 0.001$. However, constraining residuals did not demonstrate strict invariance between education groups, $\Delta\chi^2(35) = 4116.2$, $p < 0.001$, $\Delta CFI = 0.345$, $\Delta RMSEA = 0.054$, $\Delta SRMR = 0.040$.

Model interpretation

General factor

The general factor reflected overall propensity to produce correct or strategic responses and avoid committing errors of various types. Of note, this is not the same as achievement or even general cognitive ability on testing, but rather a latent construct reflected in all retained BPA metrics, and which is unique from specific factors. Likewise, specific factors reflected instrumental aspects of the NP tests from which they were derived, and subtle aspects of cognitive performance indexed by these measures. See Supplemental Tables 2a, 5b for description of BPA variables and CFA factor loadings; Supplemental Table 2b provides a description of BPA variables not retained in the model.

Letter fluency

BPA variables from letter fluency created a specific factor that was reflected in correct responses produced across time intervals. Earlier time intervals (i.e., 0–15 s and 16–30 s) had a slightly higher loading gradient, particularly for the F and A trials. Earlier time intervals also loaded higher on the general factor for the first and second trial of the task, whereas the latter time intervals loaded higher on the general factor for the S trial. There was a very slight loading gradient across time intervals and trials ($S > A > F$).

Clock drawing

BPA variables from clock drawing (command, copy, and number placement) created a specific factor derived from components of anchoring, which is considered an effective organizational strategy (Lamar et al., 2016). These variables all loaded positively on the general factor, with slightly higher loadings for the command condition and for anchor digits three and nine (across conditions). This may represent relative importance for initiating and sustaining anchoring, particularly on command.

Naming

BPA variables from naming created a specific factor that consisted of circumlocution and semantic paraphasia errors. The count of circumlocution errors loaded slightly higher on both the general and specific factor.

Digit span

BPA variables from digit span created a specific factor that appeared to separate attention (derived from digit span forward) from working memory (derived from digit span backward). Of note, the “testing the limits” variable (longest digit span with correct numbers regardless of sequence) was more reflective of this attentional component but not working memory or general ability. In addition, BPA variables related to working memory had a higher loading on the general factor.

Table 4. Measurement invariance by age group ($n = 2,363$)

Model	χ^2	df	p	CFI	TLI	RMSEA	SRMR	BIC
1. Configural	1,024.02	1,008	.356	1.00	1.00	0.00	0.02	726,172.18
2. Metric	1,051.67	1,070	.650	1.00	1.00	0.00	0.02	725,718.24
3. Scalar	1,089.04	1,097	.562	1.00	1.00	0.00	0.02	725,545.88
4. Strict	5,694.93	1,132	<.001	0.63	0.61	0.06	0.07	729,879.90
<i>Measurement invariance by sex ($n = 2,363$)</i>								
Model	χ^2	df	p	CFI	TLI	RMSEA	SRMR	BIC
1. Configural	1,026.05	1,008	.339	1.00	1.00	.00	.02	718,948.29
2. Metric	1,052.63	1,070	.642	1.00	1.00	.00	.02	718,493.27
3. Scalar	1,066.41	1,097	.741	1.00	1.00	.00	.02	718,297.32
4. Strict	1,319.31	1,132	<.001	.98	.98	.01	.03	718,278.36
<i>Measurement invariance by education group ($n = 2,363$)</i>								
Model	χ^2	df	p	CFI	TLI	RMSEA	SRMR	BIC
1. Configural	1,024.72	1,008	.350	1.0	1.0	.00	.02	725,098.18
2. Metric	1,055.30	1,070	.620	1.0	1.0	.00	.02	724,647.16
3. Scalar	1,097.42	1,097	.491	1.0	1.0	.00	.02	724,479.55
4. Strict	5,213.62	1,132	<.001	.66	.64	.06	.06	728,323.88

Note: Table displays fit for multigroup confirmatory factor analysis and then weak, strong, and strict invariance (respectively). The 60+ sample employs a Bifactor-7 model. Age groups reflect a cut point of 65. Sex reflects male or female. Education groups reflect those with a high school diploma or less versus some college or a college degree. Chi square (χ^2), degrees of freedom (df), probability (p), comparative fit index (CFI), Tucker–Lewis index (TLI), root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR), and Bayes Information Criterion (BIC).

Logical memory

BPA variables from story memory created two specific factors. The first included verbatim details that were produced during learning and recall. These counts demand a higher accuracy of recall than the traditional summary measure on the task, which allows paraphrased details. The second specific factor included conceptually related errors (i.e., new or altered details that change the meaning of target information). These errors can be considered “near misses” or “gist errors” since unlike paraphrased details, they are not acceptable for credit. Of note, the two specific factors were not significantly related, $\beta = -0.02$, 95% CI $(-0.07, 0.04)$. As expected, verbatim recall loaded positively on the general factor; however, gist errors also had a small positive loading (suggesting even “near misses” on a memory task index some element of cognitive ability).

Paired associates

BPA variables from paired associates created a specific factor that consisted of errors while learning and recalling word pairs (i.e., interference from target words not paired with the stimulus word and “gist errors” on delay, incorrect words that were conceptually related to the stimulus word). Interference errors on the third learning trial and delay were most reflective of the specific ability/tendency. Also, these variables had a negative loading on the general factor, particularly interference errors on the third learning trial.

Covariance between specific factors

We examined the covariance structure among specific factors to explore whether distinct cognitive processes co-occurred across tasks. Participants who produced more words across time intervals during letter fluency also tended to anchor on clock drawing ($\varphi = 0.11$, 95% CI $[0.05, 0.16]$), learned and recalled more verbatim details from stories ($\varphi = 0.22$, 95% CI $[0.16, 0.28]$), and had more “near misses” on story memory ($\varphi = 0.08$, 95% CI $[0.02, 0.14]$). These participants also produced fewer errors on naming ($\varphi = -0.19$, 95% CI $[-0.28, -0.10]$), and while learning/recalling word pairs ($\varphi = -0.12$, 95% CI $[-0.19, -0.05]$). There were similar associations between anchoring on clock drawing

and verbatim recall, naming errors, and errors while learning/recalling word pairs (see Supplemental Table 6). However, the specific factor derived from digit span, which largely reflected attention, was not significantly associated with other specific factors. Also, “near misses” on story memory was not associated with most other specific factors, aside from letter fluency across time intervals (as reported) and errors on naming, $\varphi = 0.11$, 95% CI $(0.02, 0.20)$. These associations suggest possible executive functioning and semantic components were distributed across the specific factors, rather than centralized in particular measures. Due to the nature of the bifactor model, these associated factors were independent from general ability. We next examined whether these latent cognitive factors predicted incipient dementia.

Structural equation models

Participants that were higher on the general factor were significantly less likely to develop all-cause dementia, OR = 0.15 (95% CI 0.12–0.86). Likewise, higher verbatim recall of story details at BPA assessment was associated with reduced odds of developing all-cause dementia, OR = 0.24 (95% CI 0.230.90). Other specific factors were not significantly related to future conversion (see Table 5). In these same models, older age and lower education were broadly associated with lower general factor scores ($\beta_{\text{age}} = -0.12$, 95% CI $[-0.17, -0.06]$; $\beta_{\text{edu}} = 0.14$, 95% CI $[0.09, 0.20]$) and reduced verbatim recall ($\beta_{\text{age}} = -0.18$, 95% CI $[-0.23, -0.14]$; $\beta_{\text{edu}} = 0.19$, 95% CI $[0.15, 0.24]$). This pattern of association was generally consistent for most specific factors, in which older participants with lower educational attainment produced fewer correct or strategic responses and committed more errors of various types. Sex differences were more nuanced, with females outperforming males on several tests including recall ($\beta = 0.10$, 95% CI $[0.05, 0.14]$) and paired associates ($\beta = -0.07$, 95% CI $[-0.13, -0.01]$), but showing more errors on naming ($\beta = 0.10$, 95% CI $[0.03, 0.17]$) and worse performance on the attention factor ($\beta = -0.07$, 95% CI $[-0.12, -0.01]$). Biological sex was not significantly associated with gist errors on story memory ($\beta = 0.03$, 95% CI $[-0.02, 0.08]$). See Supplemental Figure 1 for

Table 5. Regression table for the 60+ sample ($N = 2,363$)

Outcome	Predictor	<i>b</i>	<i>SE</i>	<i>Z</i>	<i>p</i>	<i>b*</i>	95% CI
G	Education	0.005	0.001	5.02	<.001***	0.14	[0.09, 0.20]
	Age	-0.003	0.001	-4.01	<.001***	-0.12	[-0.17, -0.06]
	Sex	0.001	0.001	0.63	.528	0.02	[-0.04, 0.07]
Dementia	G	-1.177	0.335	-3.52	<.001***	-0.10	[-0.15, -0.04]
	Education	0.008	0.001	8.43	<.001***	0.21	[0.17, 0.26]
	Age	-0.005	0.001	-6.32	<.001***	-0.16	[-0.21, -0.11]
BiF1	Sex	0.003	0.001	3.08	.002**	0.07	[0.03, 0.12]
	Education	0.004	0.001	4.91	<.001***	0.12	[0.07, 0.16]
	Age	-0.001	0.001	-0.90	.366	-0.02	[-0.07, 0.02]
BiF2	Sex	0.002	0.001	2.19	.029*	0.05	[0.01, 0.10]
	Education	0.007	0.001	7.86	<.001***	0.19	[0.15, 0.24]
	Age	-0.006	0.001	-7.62	<.001***	-0.18	[-0.23, -0.14]
BiF3	Sex	0.004	0.001	4.11	<.001***	0.10	[0.05, 0.14]
	Education	0.002	0.001	1.92	.055	0.05	[-0.00, 0.10]
	Age	0.001	0.001	1.52	.128	0.04	[-0.01, 0.09]
BiF4	Sex	0.001	0.001	1.30	.195	0.03	[-0.02, 0.08]
	Education	0.003	0.001	1.72	.085	0.07	[-0.01, 0.15]
	Age	-0.002	0.001	-2.12	.034*	-0.08	[-0.15, -0.01]
BiF5	Sex	-0.003	0.001	-2.40	.016*	-0.07	[-0.12, -0.01]
	Education	-0.005	0.001	-3.43	.001***	-0.13	[-0.21, -0.06]
	Age	0.006	0.001	4.71	<.001***	0.18	[0.11, 0.26]
BiF6	Sex	0.004	0.002	2.77	.006**	0.10	[0.03, 0.17]
	Education	-0.003	0.001	-2.45	.014*	-0.07	[-0.13, -0.02]
	Age	-0.001	0.001	-0.59	.556	-0.02	[-0.08, 0.04]
BiF7	Sex	-0.003	0.001	-2.37	.018*	-0.07	[-0.13, -0.01]
	Education	-0.440	0.340	-1.29	.196	-0.03	[-0.08, 0.02]
	Age	-0.245	0.327	-0.75	.454	-0.02	[-0.06, 0.03]
Dementia	BiF1	-0.883	0.327	-2.70	.007**	-0.07	[-0.11, -0.02]
	BiF2	0.041	0.348	0.12	.906	0.00	[-0.05, 0.05]
	BiF3	-0.139	0.385	-0.36	.717	-0.01	[-0.06, 0.04]
	BiF4	0.402	0.481	0.84	.403	0.03	[-0.04, 0.10]
	BiF5	-0.115	0.413	-0.28	.780	-0.01	[-0.07, 0.05]

Note: Table displays non-standardized and standardized effects from the Bifactor-7 model for general (G) and specific factors (BiF1-7) derived from Boston Process Approach (BPA) variables. Latent factors were predicted by educational attainment, age at BPA assessment, and sex (male or female). Future conversion to dementia (Dementia) is predicted by these same latent factors. Pathways involving the general and specific factors were assessed separately given the complexity of the model and interpretation of these factors.

visualization of demographic pathways. In sum, latent general ability and verbatim recall were the strongest predictors of incident dementia, reinforcing their potential as early cognitive markers. In addition, the association between demographic variables and latent factors in the model demonstrated better performance in general for younger participants with higher educational attainment.

Machine learning models

Models with only NP summary scores accurately predicted dementia conversion in 88.4% of participants (PR AUC = 0.96, AUC = 0.73; pooled across five imputations). Models with BPA data had 90.2% accuracy (PR AUC = 0.95, AUC = 0.74), whereas combined NP/BPA data altered classification performance to 90.1% accuracy (PR AUC = 0.96, AUC = 0.76). Most importantly, sensitivity increased to detect almost all cases of dementia conversion in models with BPA variables. Across imputations, 99.4% of converters were identified with BPA data and 99.1% with combined NP/BPA data, versus 95.9% without BPA data. There were also fewer false positives (87.6% specificity with NP data, 89.2% with BPA data, and 89.1% with combined NP/BPA data). In permutation-importance analyses, BPA variables accounted for 83.9% of total predictive contribution, representing 88.0% of the 25 most influential predictors and 80.0% of the top 50 (see Supplemental Figure 2). Among the most influential predictors were letter and semantic fluency subcomponents, instructional errors (e.g., violating instructions to not lift pen on trail-making), errors suggesting inability to complete recall tasks, perseveration

errors, isolate errors, partially correct responses (similarities), circumlocution and phonemic paraphasia errors on naming, self-corrected errors on clock drawing (command), and story gist recall.

Discussion

Quantified BPA data highlight the richness of NP testing, capturing subtle aspects of cognitive performance that are often unexamined with traditional summary scores. Across multiple measures, these results suggest summary scores are not complete indicators of the abilities that determine process and achievement on a task. As a complement to these traditional scores, constituent BPA data provide important and nuanced information on process and achievement. The novel methodology employed in this study enabled data-driven model development, accommodating the complexity and heterogeneity of the data. Our results suggest that multi-dimensional bifactor models are appropriate, with a general factor that reflects both propensity for correct/strategic responses and avoidance of errors. In addition, distinct NP tests yielded one or more specific latent factors, likely capturing test-specific variance while also reflecting nuanced cognitive features based on differences in factor loadings, dimensionality, and covariance. In structural models, we found that demographics were associated with latent factors in expected patterns, and that these same factors predicted incipient dementia. Based on the dataset, exploratory machine learning (ML) models identified 220 to 221 of the 222 participants who were later diagnosed with all-cause dementia (99.3% across imputations).

These findings have both psychometric and clinical relevance in reinforcing and informing the clinical use of NP tests. Confirmatory fit reflected the distinct cognitive abilities indexed by tasks such as digit span forward and backward. Specific errors, such as gist errors on story memory or interference during paired-associate learning, loaded onto discrete factors, consistent with prior work demonstrating their distinctness and diagnostic relevance (Loewenstein et al., 2004). Model fit also captured aspects of task process and achievement reflected in BPA data. For example, the loadings for letter fluency were not identical across time intervals and trials. While the overall range was narrow (0.38–0.48), the observed gradients could reflect how trial segments relate to underlying ability, consistent with commonly described clinical observation of task initiation and persistence. Likewise, digit span sequence recall versus number recall differentially reflected attentional versus working memory demands that would be aggregated with traditional scoring. Although modest, these differences align with standard clinical interpretation of performance.

Another noteworthy exception to conventional scoring was gist errors on story memory (i.e., new or altered details that are conceptually related to target information, but which do not qualify for credit). These “errors” were weakly but positively associated with the general factor. This suggests that gist-based errors may reflect a subtle aspect of memory *ability*, which is not captured by a summary score. Understanding “near misses” as reflective of ability is consistent with the BPA, and with extensive research on gist memory in neurodegenerative disorders. For example, previous studies suggest gist memory may serve as a compensatory or backup memory system when verbatim recall fails, which is also compromised in patients with AD compared to healthy controls (see Budson et al., 2006; Deason et al., 2012; Gallo et al., 2006).

It is also notable that in the partitioning of signal and noise, variables derived from tests associated with executive resources loaded the highest on the general factor (Marko et al., 2023; Roussel et al., 2012; Waris et al., 2017), particularly Digit Span (Backward > Forward; Gignac & Weiss, 2015). Importantly, however, almost all NP tests contributed to the general factor, and inter-factor associations suggested distributed executive and semantic processes that were influential across tests. Taken together, these findings challenge the one-test-one-domain framework, and highlight the BPA's utility in capturing the full cognitive profile needed to inform diagnosis (Ashendorf et al., 2013; Casaletto & Heaton, 2017; Harvey, 2019).

Our psychometric analyses suggested that BPA-derived factors had stable measurement properties across age, sex, and education groups in older adults. Thus, observed differences in latent variables reflect genuine performance variation rather than measurement artifacts (Borsboom, 2006; Millsap, 1997). Most of the demographic effects followed expected patterns, in which older participants with lower educational attainment produced fewer correct or strategic responses and committed more errors. However, some variables such as anchoring during clock drawing, “near misses” on story memory, and errors while learning/recalling word pairs were robust to these demographic influences. Similarly, sex-based differences were mixed, with females outperforming males on some specific factors but not on general ability.

This study reveals the feasibility of using complex models with quantified BPA data to identify individuals at risk for future cognitive decline. While traditional NP summary scores contained considerable predictive value for incipient dementia, BPA data

improved those models. ML methods were able to identify almost all cases of dementia conversion, and BPA variables were highly influential in their predictive contribution. We prioritized sensitivity in our interpretation because of the importance to early detection, and given relatively limited consequences for specificity (i.e., aside from possible unnecessary testing/monitoring). It is debatable whether the modest improvement in model performance supports BPA data collection. However, for individuals impacted, even a small improvement could have meaningful clinical value. In addition, as stated in the introduction, digital technologies will likely make such collection more feasible in clinical settings. These models could be used to generate individualized risk estimates for dementia, although external validation in multiple samples would be required before they could be considered reliable for clinical use.

Most importantly, the methods used in this study permit an interpretable structural model even with a small number of tests for each cognitive function. Based on the approximate odds ratios, a standard deviation reduction in the general factor was associated with over sixfold increased odds of converting to dementia at a later timepoint. Likewise, a standard deviation reduction in verbatim recall on story memory was associated with approximately fourfold greater odds of dementia conversion. Other variables were informative primarily through their contribution to general ability. In sum, BPA data are sensitive to subtle cognitive changes that predict incipient dementia. However, it is the combination of psychometric and ML methods used in this study that provides complementary support for the value of NP and BPA data. On their own, ML models do not clarify measurement properties or reveal what tests measure (e.g., underlying abilities translatable to daily life), and psychometric models do not prioritize predictive accuracy over other model objectives.

There are limitations to our study. While there are numerous advantages to the methodological approach, the structural models we employed do not permit survival analyses (i.e., prediction of time to dementia). Future studies could employ the models derived from this study in hierarchical time-to-event models (e.g., using similar methodology to Nichols et al., 2024), as well as alternate psychometric approaches including Multidimensional Item Response Theory and Multiple Correspondence Analysis (Greenacre, 1988; Reckase, 1997). Based on restricted ethnic and cultural diversity in the sample, we were unable to assess MI for these groups or validly assess group differences on any of the latent factors. In the future, this should be possible given FHS efforts to increase diversity with Omni 1 and Omni 2 cohorts (Andersson et al., 2019). In addition, we examined prediction of all-cause dementia; future studies should employ similar approaches with specific neurodegenerative etiologies, as different BPA variables could be sensitive to decline in each of these groups.

Retained structural models do not include most available BPA variables. This is likely due to the zero-inflated nature for many of these counts. For some tests, this excluded specific types of errors or strategic responses; for other tests, models were unable to distinguish between erroneous responses and lack of response. This study is proposed as a foundational attempt to model this type of data in a community sample of older adults. Future studies using similar methods with specific clinical samples may find different combinations of BPA variables to be informative. In addition, we conducted exploratory ML analyses in an effort to examine the relative predictive value for all BPA variables as well as NP summary scores. As noted, we did not exclude participants presenting for BPA assessment with MCI. The sensitivity to

incipient dementia with this level of impairment was poor, and most importantly, proper diagnosis of MCI requires expert clinical review, which differs from the statistical models reported here.

This study involved psychometric modeling and empirical evaluation of constituent BPA data rather than compiled measures based on BPA theory or past research (grouping errors based on type across measures, computing ratios, etc.). As discussed in Measurements and Latent Factors, there would be error in these groupings that could not be properly evaluated. Of note, our factor models contain equivalent information to some of these groupings/ratios, but with expanded information on factor configuration, processes across tests, and error. Nonetheless, it must be acknowledged that some variables excluded from models might have been retained with compiled measures. This was intended as a foundation study using only empirical, data-driven methods, intended to support future studies employing more traditional BPA theory to derive composites. Future studies should also expand on ensemble models with both BPA variables and NP summary scores. Exploratory ML analyses in this study suggest that BPA data may assist optimal dementia prediction. It may also be possible to construct a rule-list using classification and regression trees, which could reveal the particular patients for which BPA data add value (e.g., patients with similar summary scores profiles but distinct errors that predict future conversion to dementia). However, external validation will be essential to establish generalizability of these findings.

The psychometric approaches employed in this study are presented as opportunities for researchers to better understand NP data, rather than restrictions or rules to follow (Borsboom, 2006). Study findings support the use of BPA data in studies of aging and help to bridge the gap between clinical interpretation of NP tests and research utilization of these same data.

Supplementary material. The supplementary material for this article can be found at <https://doi.org/10.1017/S1355617726101921>.

Acknowledgements. None.

Funding statement. This work was supported by the Framingham Heart Study of the National Heart Lung and Blood Institute of the National Institutes of Health and Boston University School of Medicine. This project has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services, under Contract No. 75N92019D00031. This work was also supported or informed by U19 AG068753, R01-AG072654, R01-AG062109, R01AG083735, U01AG068221, K07-AG066813, and K01-AG057798 from the National Institute on Aging, U.S. Department of Veterans Affairs (VA) Merit Award (CX002400), and a VA Career Development Awards (CX002625). The authors acknowledge the Framingham Heart Study participants and researchers for their efforts over the years, without whom this study would not be possible.

Competing interests. RA is Scientific Advisor to Signant Health and NovoNordisk. The remaining authors have no conflicts to report.

References

- Amemiya, T. (1981). Qualitative response models: A survey. *Journal of Economic Literature*, 19(4), 1483–1536.
- Amini, S., Hao, B., Yang, J., Karjadi, C., Kolachalama, V. B., Au, R., & Paschalidis, I. C. (2024). Prediction of Alzheimer's disease progression within 6 years using speech: A novel approach leveraging language models. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 20(8), 5262–5270.
- Andersson, C., Johnson, A. D., Benjamin, E. J., Levy, D., & Vasan, R. S. (2019). 70-year legacy of the Framingham heart study. *Nature Reviews. Cardiology*, 16(11), 687–698.
- Ashendorf, L., Swenson, R., & Libon, D. (2013). *The Boston process approach to neuropsychological assessment: A practitioner's guide*. (pp. xxviii–431). Oxford University Press.
- Au, R., Bose, N., Imam, F., & Kolachalama, V. B. (2025). Technological advances enabling an enhanced understanding of early Alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 21(11), e70883.
- Au, R., & Devine, S. (2013). Implementing the Boston Process Approach in the Framingham heart study: It is possible. In *The Boston process approach to neuropsychological assessment: a practitioner's guide* (pp. 380–406). Oxford University Press.
- Au, R., Popp, Z. T., Low, S., Hwang, P. H., De Anda-Duran, I., Li, S., Rahman, S., Ding, H., Igwe, A., Karjadi, C., Ang, T. F. A., Devine, S. A., Gurnani, A. S., Mez, J. B., Farrer, L. A., Sunderaraman, P., Lin, H., & Kolachalama, V. B. (2023). Digital is the new blood: Enabling the present and the future. *Alzheimer's & Dementia*, 19, e078220. <https://doi.org/10.1002/alz.078220>.
- Au, R., Seshadri, S., Wolf, P. A., Elias, M., Elias, P., Sullivan, L., Beiser, A., & D'Agostino, R. B. (2004). New norms for a new generation: Cognitive performance in the Framingham offspring cohort. *Experimental Aging Research*, 30(4), 333–358.
- Bartlett, M. S. (1951). The effect of standardization on a χ^2 approximation in factor analysis. *Biometrika*, 3(4), 337–344.
- Bélanger, S., & Belleville, S. (2009). Semantic inhibition impairment in mild cognitive impairment: A distinctive feature of upcoming cognitive decline? *Neuropsychology*, 23(5), 592–606.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107(2), 238–246.
- Bettcher, B. M., Giovannetti, T., MacMullen, L., & Libon, D. J. (2008). Error detection and correction patterns in dementia: A breakdown of error monitoring processes and their neuropsychological correlates. *Journal of the International Neuropsychological Society*, 14(2), 199–208.
- Binaco, R., Calzaretto, N., Epifano, J., McGuire, S., Umer, M., Emrani, S., Wasserman, V., Libon, D. J., & Polikar, R. (2020). Machine learning analysis of digital clock drawing test performance for differential classification of mild cognitive impairment subtypes versus Alzheimer's disease. *Journal of the International Neuropsychological Society*, 26(7), 690–700.
- Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, 71(3), 425–440.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Sage.
- Budson, A. E., Todman, R. W., & Schacter, D. L. (2006). Gist memory in Alzheimer's disease: Evidence from categorized pictures. *Neuropsychology*, 20(1), 113–122.
- Casaletto, K. B., & Heaton, R. K. (2017). Neuropsychological assessment: Past and future. *Journal of the International Neuropsychological Society: JINS*, 23(9–10), 778–790.
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 464–504.
- Cummings, J., Zhou, Y., Lee, G., Zhong, K., Fonseca, J., & Cheng, F. (2024). Alzheimer's disease drug development pipeline: 2024. *Alzheimer's & Dementia*, 10(2), e12465.
- Davoudi, A., Dion, C., Amini, S., Tighe, P. J., Price, C. C., Libon, D. J., & Rashidi, P. (2021). Classifying non-dementia and Alzheimer's disease/vascular dementia patients using kinematic, time-based, and visuospatial parameters: The digital clock drawing test. *Journal of Alzheimer's Disease*, 82(1), 47–57.
- Deason, R. G., Hussey, E. P., Budson, A. E., & Ally, B. A. (2012). Gist-based conceptual processing of pictures remains intact in patients with amnesic mild cognitive impairment. *Neuropsychology*, 26(2), 202–208.
- Ding, H., Lister, A., Karjadi, C., Au, R., Lin, H., Bischoff, B., & Hwang, P. H. (2024). Detection of mild cognitive impairment from non-semantic, acoustic voice features: The Framingham heart study. *JMIR Aging*, 7, e51126.
- Dion, C., Frank, B. E., Crowley, S. J., Hize, L. P., Rodriguez, K., Tanner, J. J., Libon, D. J., & Price, C. C. (2021). Parkinson's disease cognitive phenotypes show unique clock drawing features when measured with digital technology. *Journal of Parkinson's Disease*, 11(2), 779–791.

- Emond, E. J., & Mason, D. W. (2002). A new rank correlation coefficient with application to the consensus ranking problem. *Journal of Multi-Criteria Decision Analysis*, 11(1), 17–28.
- Eppig, J., Wambach, D., Nieves, C., Price, C. C., Lamar, M., Delano-Wood, L., Giovannetti, T., Bettcher, B. M., Penney, D. L., Swenson, R., Lippa, C., Kabasakalian, A., Bondi, M. W., & Libon, D. J. (2012). Dysexecutive functioning in mild cognitive impairment: Derailment in temporal gradients. *Journal of the International Neuropsychological Society*, 18(1), 20–28.
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3), 272–299.
- Ferretti, M. T., Ding, H., Au, R., Liu, C., Devine, S., Auerbach, S., Mez, J., Gurnani, A., Liu, Y., Santucci, A., & Ang, T. F. A. (2024). Maximizing utility of neuropsychological measures in sex-specific predictive models of incident Alzheimer's disease in the Framingham heart study. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 20(2), 1112–1122.
- Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research: JMLR*, 20, 177.
- Frank, B., Bandyopadhyay, S., Dion, C., Formanski, E., Matusz, E., Penney, D., Davis, R., O'Connor, M. K., Au, R., Amini, S., Rashidi, P., Tighe, P., Libon, D. J., & Price, C. C. (2024). A network analysis of digital clock drawing for command and copy conditions. (Sage CA: Los Angeles, CA). *Assessment*, 32, 119–129.
- Gallo, D. A., Shahid, K. R., Olson, M. A., Solomon, T. M., Schacter, D. L., & Budson, A. E. (2006). Overdependence on degraded gist memory in Alzheimer's disease. *Neuropsychology*, 20(6), 625–632.
- Gignac, G. E., & Weiss, L. G. (2015). Digit span is (mostly) related linearly to general intelligence: Every extra bit of span counts. *Psychological Assessment*, 27(4), 1312–1323.
- Greenacre, M. J. (1988). Correspondence analysis of multivariate categorical data by weighted least-squares. *Biometrika*, 75(3), 457–467.
- Gupta, A., Preis, S. R., Beiser, A., Devine, S., Hanke, L., Seshadri, S., Wolf, P. A., & Au, R. (2015). Midlife cardiovascular risk impacts memory function: The Framingham offspring study. *Alzheimer Disease and Associated Disorders*, 29(2), 117–123.
- Guttman, L. (1955). The determinacy of factor score matrices with implications for five other basic problems of common-factor theory. *British Journal of Statistical Psychology*, 8, 65–81.
- Hankee, L. D., Preis, S. R., Beiser, A. S., Devine, S. A., Liu, Y., Seshadri, S., Wolf, P. A., & Au, R. (2013). Qualitative neuropsychological measures: Normative data on executive functioning tests from the Framingham offspring study. *Experimental Aging Research*, 39(5), 515–535.
- Harvey, P. D. (2019). Domains of cognition and their assessment. *Dialogues in Clinical Neuroscience*, 21(3), 227–237.
- Holzinger, K. J., & Swineford, F. (1937). The Bi-factor method. *Psychometrika*, 2(1), 41–54.
- Hooper, E. (1983). *Hooper visual organization test (VOT)*. Western Psychological Services.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179–185.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55.
- Jak, A. J., Bondi, M. W., Delano-Wood, L., Wierenga, C., Corey-Bloom, J., Salmon, D. P., & Delis, D. C. (2009). Quantification of five neuropsychological approaches to defining mild cognitive impairment. *The American Journal of Geriatric Psychiatry: Official Journal of the American Association for Geriatric Psychiatry*, 17(5), 368–375.
- Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39(1), 31–36.
- Kaplan, E. (1988). A process approach to neuropsychological assessment. In *Clinical neuropsychology and brain function: Research, measurement, and practice* (pp. 127–167). American Psychological Association. <https://doi.org/10.1037/10063-004>.
- Kaplan, E. (1990). The process approach to neuropsychological assessment of psychiatric patients. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 2(1), 72–87.
- Kaplan, E., Goodglass, H., & Weintraub, S. (1983). *The Boston Naming Test*. Lea & Fibiger.
- Kolo, F. B., Lu, S., Beiser, A. S., Francis, L., Melo van Lent, D., Gireud-Goss, M., Himali, J. J., Ghosh, S., Ting, S. J., Banerjee, A., Kowalczyk, T. L., Zucker, J. M., Heard-Costa, N., Wood, C. G., Maillard, P., Fletcher, E. M., DeCarli, C. S., Welling, D. B., Kujawa, S. G., & Seshadri, S. (2025). Hearing loss, brain structure, cognition, and dementia risk in the Framingham heart study. *JAMA Network Open*, 8(11), e2539209.
- Lamar, M., Ajilore, O., Leow, A., Charlton, R., Cohen, J., GadElkarim, J., Yang, S., Zhang, A., Davis, R., Penney, D., Libon, D. J., & Kumar, A. (2016). Cognitive and connectome properties detectable through individual differences in graphomotor organization. *Neuropsychologia*, 85, 301–309.
- Lamar, M., Podell, K., Carew, T. G., Cloud, B. S., Resh, R., Kennedy, C., Goldberg, E., Kaplan, E., & Libon, D. J. (1997). Perseverative behavior in Alzheimer's disease and subcortical ischemic vascular dementia. *Neuropsychology*, 11(4), 523–534.
- Lamar, M., Price, C. C., Giovannetti, T., Swenson, R., & Libon, D. J. (2010). The dysexecutive syndrome associated with ischaemic vascular disease and related subcortical neuropathology: A Boston process approach. *Behavioural Neurology*, 22(1–2), 53–62.
- Libon, D. J., Bondi, M. W., Price, C. C., Lamar, M., Eppig, J., Wambach, D. M., Nieves, C., Delano-Wood, L., Giovannetti, T., Lippa, C., Kabasakalian, A., Cosentino, S., Swenson, R., & Penney, D. L. (2011). Verbal serial list learning in mild cognitive impairment: A profile analysis of interference, forgetting, and errors. *Journal of the International Neuropsychological Society*, 17(5), 905–914.
- Libon, D. J., Lamar, M., Price, C. C., Jefferson, A. L., Swenson, R., & Au, R. (2018). Neuropsychological evaluation for vascular dementia. In G. E. Smith & S. T. Farias (Eds.), *APA handbook of dementia* (pp. 325–341). American Psychological Association. <https://doi.org/10.1037/0000076-016>.
- Libon, D. J., Preis, S. R., Beiser, A. S., Devine, S., Seshadri, S., Wolf, P. A., DeCarli, C., & Au, R. (2015). Verbal memory and brain aging: An exploratory analysis of the role of error responses in the Framingham study. *American Journal of Alzheimer's Disease and Other Dementias*, 30(6), 622–628.
- Libon, D. J., Swenson, R., Lamar, M., Price, C. C., Baliga, G., Pascual-Leone, A., Au, R., Cosentino, S., Andersen, S. L., & Loewenstein, D. (2022). The Boston process approach and digital neuropsychological assessment: Past research and future directions. *Journal of Alzheimer's Disease*, 87(4), 1419–1432.
- Libon, D. J., Swenson, R., Langford, D. T., Cosentino, S., Price, C. C., Lamar, M., Emrani, S., Au, R., Andersen, S., Chen, M. H., Ashendorf, L., & Thompson, L. (2025). Precision neurocognition: An emerging diagnostic paradigm leveraging digital cognitive assessment technology. *Journal of Alzheimer's Disease: JAD*, 108(1_suppl), 13872877251325725.
- Libon, D. J., Swenson, R., Price, C. C., Lamar, M., Cosentino, S., Bezdicek, O., Kling, M. A., Tobyn, S., Jannati, A., Banks, R., & Pascual-Leone, A. (2024). Digital assessment of cognition in neurodegenerative disease: A data driven approach leveraging artificial intelligence. *Frontiers in Psychology*, 15, 1415629.
- Libon, D. J., Swenson, R., Tobyn, S., Jannati, A., Schulman, D., Price, C. C., Lamar, M., & Pascual-Leone, A. (2024). Dysexecutive difficulty and subtle everyday functional disabilities: The digital trail making test. *Frontiers in Neurology*, 15, 1354647.
- Lim, S., & Jahng, S. (2019). Determining the number of factors using parallel analysis and its recent variants. *Psychological Methods*, 24(4), 452–467.
- Loewenstein, D. A., Acevedo, A., Luis, C., Crum, T., Barker, W. W., & Duara, R. (2004). Semantic interference deficits and the detection of mild Alzheimer's disease and mild cognitive impairment without dementia. *Journal of the International Neuropsychological Society*, 10(1), 91–100.
- Loewenstein, D. A., Acevedo, A., Amarilis, A., Joscelyn, Martinez, Glenda, & Duara, R. (2007). The use of amnesic and nonamnesic composite measures at different thresholds in the neuropsychological diagnosis of MCI*. *Journal of Clinical & Experimental Neuropsychology*, 29(3), 300–307.
- Marko, M., Michalko, D., Dragašek, J., Vančová, Z., Jarčuškova, D., & Riečanský, I. (2023). Assessment of automatic and controlled retrieval using verbal fluency tasks. *Assessment*, 30(7), 2198–2211.
- McDonald, R. P., & Bolt, D. M. (1998). The determinacy of variables in structural equation models. *Multivariate Behavioral Research*, 33(3), 385–401.

- McNeish, D. (2023). Psychometric properties of sum scores and factor scores differ even when their correlation is 0.98: A response to Widaman and Revelle. *Behavior Research Methods*, 55(8), 4269–4290.
- Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods*, 2(3), 248–260.
- Müller, S., Herde, L., Preische, O., Zeller, A., Heymann, P., Robens, S., Elbing, U., & Laske, C. (2019). Diagnostic value of digital clock drawing test in comparison with CERAD neuropsychological battery total score for discrimination of patients in the early course of Alzheimer's disease from healthy individuals. *Scientific Reports*, 9, 3543.
- Nichols, E., Aslanyan, V., Adrien, T. V., Andrews, R. M., Fardo, D. W., Gavett, B. E., Paterson, T. S. E., Turney, I. C., Young, C. B., Uanhoro, J. O., Gross, A. L., & Initiative, for the A. D. N. (2024). Measurement error and methodologic issues in analyses of the proportion of variance explained in cognition. *Neuropsychology Review*, 35(4), 731–744.
- Nishtala, A., Preis, S. R., Beiser, A., Devine, S., Hanke, L., Seshadri, S., Wolf, P. A., & Au, R. (2014). Midlife cardiovascular risk impacts executive function: Framingham offspring study. *Alzheimer Disease and Associated Disorders*, 28(1), 16–22.
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review: DR*, 41, 71–90.
- Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21(1), 25–36.
- Revelle, W. (2024). *psych: Procedures for psychological, psychometric, and personality research* (Version 2.4.6.26) [Computer software]. <https://cran.r-project.org/web/packages/psych/index.html>.
- Revelle, W., & Rocklin, T. (1979). Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research*, 14(4), 403–414.
- Rigdon, E. E., Becker, J.-M., & Sarstedt, M. (2019). Factor indeterminacy as metrological uncertainty: Implications for advancing psychological measurement. *Multivariate Behavioral Research*, 54(3), 429–443.
- Rosseel, Y., Oberski, D. L., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M., & Barendse, M. (2014). *lavaan: Latent variable analysis* [Computer software].
- Roussel, M., Dujardin, K., Hénon, H., & Godefroy, O. (2012). Is the frontal dysexecutive syndrome due to a working memory deficit? Evidence from patients with stroke. *Brain*, 135(7), 2192–2201.
- Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS one*, 10(3), e0118432.
- Salmon, D. P., & Bondi, M. W. (2009). Neuropsychological assessment of dementia. *Annual Review of Psychology*, 60, 257–282.
- Satizabal, C. L., Beiser, A. S., Chouraki, V., Chêne, G., Dufouil, C., & Seshadri, S. (2016). Incidence of dementia over three decades in the framingham heart study. *New England Journal of Medicine*, 374(6), 523–532.
- Spreen, O., & Strauss, E. (1991). *A compendium of neuropsychological tests: Administration, norms, and commentary*. (pp. xv–442). Oxford University Press.
- Tavabi, N., Stück, D., Signorini, A., Karjadi, C., Al Hanai, T., Sandoval, M., Lemke, C., Glass, J., Hardy, S., Lavalley, M., Wasserman, B., Ang, T. F. A., Nowak, C. M., Kaikaryam, R., Foschini, L., & Au, R. (2022). Cognitive digital biomarkers from automated transcription of spoken language. *The Journal of Prevention of Alzheimer's Disease*, 9(4), 791–800.
- Tran, U. S., & Formann, A. K. (2009). Performance of parallel analysis in retrieving unidimensionality in the presence of binary data. *Educational & Psychological Measurement*, 69(1), 50–61.
- Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41(3), 321–327.
- Waller, N. G. (2023). Breaking our silence on factor score indeterminacy. *Journal of Educational and Behavioral Statistics*, 48(2), 244–261.
- Waris, O., Soveri, A., Ahti, M., Hoffing, R. C., Ventus, D., Jaeggi, S. M., Seitz, A. R., & Laine, M. (2017). A latent factor analysis of working memory measures using large-scale data. *Frontiers in Psychology*, 8, 1062.
- Wechsler, D. (1945). A standardized memory scale for clinical use. *Journal of Psychology*, 19(1), 87–95.
- Wechsler, D. (1955). *Manual for the Wechsler Adult Intelligence Scale*. Psychological Corporation.
- Wechsler, D. (2008). *Wechsler Adult Intelligence Scale—Fourth Edition (WAIS-IV)*. Pearson.
- Young, C. B., Smith, V., Karjadi, C., Grogan, S.-M., Ang, T. F. A., Insel, P. S., Henderson, V. W., Sumner, M., Poston, K. L., Au, R., & Mormino, E. C. (2024). Speech patterns during memory recall relates to early tau burden across adulthood. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 20(4), 2552–2563.
- Zheng, X., Zhang, W., Wang, X., Li, R., Liu, M., Xu, F., Li, Y., Zheng, J., & Nie, Z. (2021). Extended application of digital clock drawing test in the evaluation of Alzheimer's disease based on artificial intelligence and the neural basis. *Curr Alzheimer Res*, 18(14), 1127–1139.