

Cognitive & Behavioral Assessment

Measuring longitudinal cognition: Individual tests versus composites

Erin M. Jonaitis^{a,*}, Rebecca L. Kosciak^a, Lindsay R. Clark^{a,b,c}, Yue Ma^c, Tobey J. Betthausen^c, Sara E. Berman^c, Samantha L. Allison^{b,c}, Kimberly D. Mueller^{a,c,d}, Bruce P. Hermann^e, Carol A. Van Hulle^c, Bradley T. Christian^{f,g}, Barbara B. Bendlin^c, Kaj Blennow^{h,i}, Henrik Zetterberg^{h,i,j,k}, Cynthia M. Carlsson^{b,a,c}, Sanjay Asthana^{b,c}, Sterling C. Johnson^{b,a,c}

^aWisconsin Alzheimer's Institute, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA

^bGeriatric Research Education and Clinical Center, William S. Middleton Memorial Veterans Hospital, Madison WI, USA

^cWisconsin Alzheimer's Disease Research Center, University of Wisconsin-Madison School of Medicine and Public Health, Madison, WI, USA

^dDepartment of Communication Sciences and Disorders, University of Wisconsin, Madison, WI, USA

^eDepartment of Neurology, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA

^fDepartment of Medical Physics, University of Wisconsin-Madison School of Medicine and Public Health, Madison, WI, USA

^gDepartment of Psychiatry, University of Wisconsin-Madison School of Medicine and Public Health, Madison, WI, USA

^hDepartment of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology, the Sahlgrenska Academy at the University of Gothenburg, Mölndal, Sweden

ⁱClinical Neurochemistry Laboratory, Sahlgrenska University Hospital, Mölndal, Sweden

^jDepartment of Neurodegenerative Disease, UCL Institute of Neurology, Queen Square, London, UK

^kUK Dementia Research Institute at UCL, London, UK

Abstract

Introduction: Longitudinal cohort studies of cognitive aging must confront several sources of within-person variability in scores. In this article, we compare several neuropsychological measures in terms of longitudinal error variance and relationships with biomarker-assessed brain amyloidosis (A β).

Methods: Analyses used data from the Wisconsin Registry for Alzheimer's Prevention. We quantified within-person longitudinal variability and age-related trajectories for several global and domain-specific composites and their constituent scores. For a subset with cerebrospinal fluid or amyloid positron emission tomography measures, we examined how A β modified cognitive trajectories.

Results: Global and theoretically derived composites exhibited lower intraindividual variability and stronger age $\times$ A β interactions than did empirically derived composites or raw scores from single tests. For example, the theoretical executive function outperformed other executive function scores on both metrics.

Discussion: These results reinforce the need for careful selection of cognitive outcomes in study design, and support the emerging consensus favoring composites over single-test measures.

© 2018 The Authors. Published by Elsevier Inc. on behalf of the Alzheimer's Association. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Keywords:

Biostatistics; Longitudinal data analysis; Cognitive aging; Neuropsychological tests; Composite scores; Intraindividual variability

E.M.J. was employed by Frontier Science from 2015 to 2017 and received funding for travel from Frontier Science and from Actelion Pharmaceuticals during that time. Her spouse is employed by Epic Systems Corporation and owns both stock and stock appreciation rights in the company. K.B. has served as a consultant or at advisory boards for Alzheon, BioArctic, Biogen, Eli Lilly, Fujirebio Europe, IBL International, Merck, Novartis, Pfizer, and Roche Diagnostics, and is a cofounder of Brain Biomarker Solutions in Gothenburg AB, a GU Venture-based platform company at the University of Gothenburg. H.Z. has served at scientific advisory boards for Roche Diagnostics, Eli Lilly, Wave, Samumed, and CogRx, has received travel support from Teva, and is a cofounder

of Brain Biomarker Solutions in Gothenburg AB, a GU Ventures-based platform company at the University of Gothenburg. C.M.C. is the site PI for a grant funded jointly by NIH and Lilly (the A4 Study) and has grant funding from NIH, the VA, Bader Philanthropies, and the RECALL Foundation. S.A. reports grants from NIA/NIH during the conduct of the study. S.C.J. has served on an advisory board for Roche. Aside from other external funding listed in the acknowledgments, no other authors have disclosures to report.

*Corresponding author. Tel.: +(608) 262-1888; Fax: +(608) 265-9122. E-mail address: jonaitis@wisc.edu

1. Introduction

Understanding individual longitudinal cognitive change requires parsing multiple sources of variability in scores. In a longitudinal observational study, consistent decline may indicate true change, whereas a succession of rises and falls may not. However, true decline may be difficult to detect when changes are subtle and fluctuations over time are large—as in the beginning stages of a dementing disorder such as Alzheimer's disease (AD), where someone may meet criteria for mild cognitive impairment (MCI) at one visit but not the next [1]. Seeking measures with high test-retest reliability may not solve the problem, as the most stable tests may not be sensitive to early change. A more subtle criterion that directly assesses longitudinal variability is the intraindividual standard deviation (IISD) over repeated assessments [2]. Individuals with larger IISD may be at higher risk of subsequent dementia [1–3] or other impairment [4]; however, high IISD values in stable normal samples may be inflated by measurement error. Strategies for reducing error are necessary for understanding early cognitive decline.

To understand variability across tests and time, longitudinal studies of cognition typically include comprehensive cognitive batteries assessing many domains [5,6]. Separate analysis of each outcome without considering familywise type I error risks spurious or irreproducible findings [7]. Alternatively, to reduce multiplicity, we can average individual tests into composite scores, as in, for example, the pre-clinical Alzheimer's cognitive composite (PACC), which combines scores from tests of memory and executive function [8]. Such composite scores have attracted attention as sensitive indicators of early cognitive change [9], and the FDA has indicated openness to cognitive composite endpoints for anti-AD drug trials [10]. Several approaches to devising composites have been proposed, including the data-driven approach, in which empirical data reduction techniques such as factor analysis are used to combine scores that tend to covary [11]; the theory-driven approach, in which established neuropsychological theories are used to combine scores within a single cognitive domain [12]; and the global approach, as in the PACC, in which representative tests from multiple domains are combined in a theory-driven way to estimate overall cognitive performance [8,13]. In developing composites, reliability and validity must be considered in tandem, ensuring the composite reflects the construct of interest—a reduction in error variance must not come at the cost of a weakened relationship to the criterion [14]. If this is achieved, composite scores can limit type I error and reduce error variance, improving statistical power.

We assessed the suitability of several cognitive tests and composites for identifying cognitive change in the context of an ongoing longitudinal study of middle-aged and older adults. We aimed to (1) identify which measures have the lowest IISD, after adjusting for known sources of cognitive

variability, and (2) examine the criterion validity of each measure by assessing its association with age and with amyloid-accelerated decline during late middle age.

2. Methods

2.1. Participants

Analyses used longitudinal neuropsychological data from participants in the Wisconsin Registry for Alzheimer's Prevention (WRAP), who are cognitively unimpaired at the baseline. Only visits with complete data were included. Participants having fewer than two complete visits ($N = 397$) or reporting a baseline neurological diagnosis ($N = 43$) were excluded. In addition, to ensure our measure of longitudinal inconsistency was not inflated by the presence of clinically significant decline, we also excluded participants who were diagnosed with MCI or dementia at any visit ($N = 52$). The effect of this exclusion criterion was examined in a sensitivity analysis (Section 2.4.5). After exclusions, this standardizing sample included data from 1063 participants with 2–5 visits (mean interval = 2.51 years). Participant characteristics are summarized in Table 1.

Full-sample validity analyses compared age effects across measures. Additional validity analyses used a subset with cerebrospinal fluid (CSF) and/or [^{11}C]Pittsburgh compound B (PiB)-labeled positron emission tomography images, enabling in vivo estimates of amyloid burden ($N = 226$). To ensure the widest range of amyloidosis, this biomarker sample included 11 additional participants with MCI or dementia who had available amyloid estimates, but had been excluded from the standardizing sample. The effect of these participants on results was examined in a sensitivity analysis (Section 2.4.5).

Procedures were performed in compliance with ethical standards for human subjects research, and all participants provided informed consent.

2.2. Assessments

Participants in WRAP complete a comprehensive cognitive battery described in full elsewhere [5]. Cognitive tests incorporated in the current analyses include the Rey Auditory-Verbal Learning Test (AVLT) [15]; the Logical Memory subtest of the Wechsler Memory Scale—Revised (LM) [16]; the Brief Visuospatial Memory Test—Revised (BVMT) [17]; the Stroop test, Color–Word Interference (STROOP) [18]; the Trail Making Test, parts A and B (TMT-A and TMT-B) [19]; the Digit Symbol subtest of the Wechsler Adult Intelligence Scale—Revised (DIGSYM) [20]; the Controlled Oral Word Association Test, CFL version (CFL) [21]; and the Mini-Mental State Exam (MMSE) [22]. We quantified baseline literacy using the Reading subtest of the Wide Range Achievement Test—Third Edition [23].

Table 1
Demographic characteristics of the WRAP sample

Sample characteristic	Cognitively unimpaired sample	Biomarker subsample	Excluded from standardization sample
N	1063	226	492
Age at WRAP recruitment, y, mean (SD)	53.9 (6.5)	54.8 (6.4)	54.9 (7)
Age at first visit selected, y, mean (SD)	58.2 (6.4)	58.7 (6.1)	–
Number of study visits included, median (range)	3 (2–5)	4 (2–5)	–
Sex, male, N (%)	322 (30%)	74 (33%)	137 (28%)
Education, some college or less, N (%)	399 (38%)	74 (33%)	252 (52%)
White/Caucasian	1014 (95%)	214 (95%)	360 (73%)
Black/African American	29 (3%)	8 (4%)	95 (19%)
Spanish/Hispanic	8 (1%)	1 (0%)	30 (6%)
American Indian/Native American	9 (1%)	2 (1%)	5 (1%)
Asian	3 (0%)	1 (0%)	1 (0%)
Parental history of AD, N (%)	772 (73%)	168 (74%)	357 (73%)
WRAT-3 reading standard score, median (range)	107 (66–120)	109 (66–119)	103 (45–120)
MMSE total, median (range)	30 (23–30)	30 (26–30)	30 (25–30)
Amyloid PET data, N (%)	–	206 (91%)	–
CSF amyloid data, N (%)	–	128 (57%)	–
Amyloid positive, N (%)	–	58 (26%)	–
Too few visits	–	–	397 (81%)
Baseline neuro dx	–	–	43 (9%)
Clin dx	–	–	52 (11%)

Abbreviations: AD, Alzheimer's disease; WRAP, Wisconsin Registry for Alzheimer's Prevention; PET, positron emission tomography; CSF, cerebrospinal fluid; MMSE, Mini-Mental State Exam; WRAT, Wide Range Achievement Test.

2.3. Biomarker methods

Methods for processing CSF are described in full elsewhere [24]. Briefly, 22 mL of CSF were removed from the L3–L4 or L4–L5 vertebral interspace for each participant. These samples were processed at the Clinical Neurochemistry Laboratory at the Sahlgrenska Academy of the University of Gothenburg, Sweden. Samples were sent in batches at two time points and analyzed using commercially available enzyme-linked immunosorbent assay methods. CSF samples were assayed for A β ₄₂ and A β ₄₀ and corrected for batch as previously described [24]. 128 participants in the present study had available CSF A β ₄₂ and/or A β ₄₀.

206 participants underwent 70-minute dynamic [¹¹C]PiB positron emission tomography scans (Siemens EXACT HR+) initiated with bolus injection (nominal 555 MBq). [¹¹C]PiB radiochemical synthesis, positron emission tomography data acquisition, image processing and quantification have been described in depth previously [25]. The primary measure was average cortical [¹¹C]PiB distribution volume ratio (reference Logan graphical analysis, cerebellum gray matter reference region, $\bar{k}_2 = 0.149 \text{ min}^{-1}$ [26,27]) across eight bilateral regions of interest (angular, anterior, and posterior cingulate, medial orbitofrontal, supramarginal, middle, and superior temporal gyri, and precuneus) [28].

2.4. Statistical methods

2.4.1. Composite measures

We considered five composites based on previous factor analyses of the WRAP battery [11,29], representing

immediate learning (EMP-IMM-LRN); delayed recall (EMP-DEL-REC); executive function (EMP-EXEC-FN); story recall (EMP-LM); and visuospatial learning (EMP-BVMT) (Table 2, columns 1–5). While item inclusion in the factor analysis was guided by theoretical perspectives on cognitive decline, the loadings and factor structure were data-driven; thus we refer to these as empirical composites (EMP). Although the cohort has grown since the first factor analysis, approximately 90 percent of the standardizing sample was in the earlier sample, and the baseline demographic characteristics of the overlapping samples were similar (Supplementary Table 1). Because some tests of interest were first administered at visit 2, the average age of sample members at the first visit included in these analyses is about 4 years older than the average baseline age reported elsewhere [5]. However, the factorial invariance by age noted in the original analysis justifies assuming that the factor structure remains a reasonable fit [11].

We also considered several theoretically derived composites (THEO). Three domain-specific theoretical composites, previously used in WRAP, represent immediate learning (THEO-IMM-LRN), delayed recall (THEO-DEL-REC), and executive function (THEO-EXEC-FN) [24] (Table 2, columns 6–8). We also considered five global composites (Table 2, columns 9–13), including the global preclinical Alzheimer's composite (PACC4-MMSE) [8]; a three-test PACC version omitting MMSE, due to its limited sensitivity in middle-aged healthy samples (PACC3) [30]; and a PACC version replacing MMSE with the CFL (PACC4-CFL) [31]. Furthermore, because one PACC test, DIGSYM, is not available in the National Alzheimer's Coordinating Center

Table 2
Thirteen composite scores (columns) and the twelve raw test scores contributing to each (rows)

Raw scores	EMP-IMM-LRN	EMP-DEL-REC	EMP-LM	EMP-BVMT	EMP-FN	THEO-EXEC-FN	THEO-DEL-REC	THEO-EXEC-FN	PACC4-MMSE	PACC3-CFL	PACC4-TMTB	PACC3-TMTB
Rey AVLT Total	X*	X*	-	-	-	-	-	-	X	X	X	X
Rey AVLT Delayed	-	X	-	-	-	-	X	-	-	-	-	-
WMS-R Logical Memory-I	-	-	X	-	-	-	-	-	-	-	-	-
WMS-R Logical Memory-II	-	-	X	-	-	-	X	-	X	X	X	X
BVMT-R Total	-	-	-	X	-	-	-	-	-	-	-	-
BVMT-R Delayed	-	-	-	X	-	-	X	-	-	-	-	-
Stroop Color-Word	-	-	-	-	X	-	-	X	-	-	-	-
TMT Part A	-	-	-	-	X	-	-	X	-	-	-	-
TMT Part B	-	-	-	-	X	-	-	X	-	-	X	X
WAIS-R Digit Symbol	-	-	-	-	-	-	-	X	X	X	-	-
COWAT C.F.L.	-	-	-	-	-	-	-	-	-	X	-	-
Mini-Mental State Exam	-	-	-	-	-	-	-	-	X	-	-	-

Abbreviations: AVLT, Auditory-Verbal Learning Test; BVMT-R, Brief Visuospatial Memory Test—Revised; COWAT, Controlled Oral Word Association Test; DEL-REC, delayed recall; EMP, empirical composites; EXEC-FN, executive function; IMM-LRN, immediate learning; LM, Logical Memory; MMSE, Mini-Mental State; PACC, preclinical Alzheimer's cognitive composite; THEO, theoretical composites; TMT, Trail Making Test; WMS-R, Wechsler Memory Scale—Revised.

NOTE. X in a cell indicates that the test represented in that row contributed to that column's composite. Empirical composite inputs (columns 1–5) were weighted according to the factor analysis on which they were based, as described by Kosciak et al. [29]. Theoretical composites (columns 6–13) were computed using equal weights.

*Empirical factor analysis suggested alternate division of immediate and delayed portions of AVLT. EMP-IMM-LRN includes information from AVLT immediate trials 1 and 2; EMP-DEL-REC includes information from immediate trials 3–5 and delayed recall.

Uniform Data Set, Third Edition [6], we included two experimental versions of the PACC4 substituting TMT-B for DIGSYM, both with (PACC4-TMTB) and without (PACC3-TMTB) MMSE. Finally, we considered individual tests contributing to each composite.

To compute composites, we first standardized all scores (mean = 0, SD = 1). Where lower scores indicated better performance (TMT-A, TMT-B), scores were multiplied by -1 . Each composite was created as an average of selected standardized raw scores (Table 2), with weighting scheme varying by composite type. Empirical composite inputs were weighted according to the factor analysis on which they were based, as described by Kosciak et al. [29]. Domain-specific and global composites were unweighted averages of their components. All composites were then re-standardized to a mean of 0 and a standard deviation of 1.

2.4.2. Convergent and discriminant validity

We explored Spearman intercorrelations among all raw and composite scores. To explore the domain structure of the theoretical composites in a systematic way, we constructed a correlation matrix of constituent raw scores (similar to a multitrait-multimethod matrix [32]). Reliability estimates (diagonal) were calculated using intraclass correlation; between-outcome estimates (off-diagonal) were calculated using the repeated measures correlation, which adjusts for between-subjects performance differences [33,34].

2.4.3. Intraindividual longitudinal standard deviation

We estimated the longitudinal inconsistency of each outcome after factoring out known sources of variability. To do this, we constructed random-slopes models of each outcome, controlling for age, sex, education, literacy, and number of prior exposures to the battery, and output the residuals, such that the score for each variable at each person-visit represented the deviation from its predicted value given the covariates. For each subject and outcome, we then calculated the IISD of these residuals as a measure of inconsistency [35]. This provided a subjectwise estimate of the amount of longitudinal within-person variability not associated with known covariates.

2.4.4. Criterion validity

Criterion validity was assessed by exploring relationships between each outcome, age, and (in the biomarker subsample) A β status. To examine age-related change across outcomes, we plotted 95% CIs of the $\hat{\beta}_{age}$ terms obtained from linear mixed models of each outcome controlling for covariates.

Primary subsample analyses treated A β as a binary variable, with 1 representing suprathreshold levels of PiB, CSF-A β_{42} , or CSF-A $\beta_{42/40}$, and 0 representing subthreshold values on each available marker. The processes for determining these thresholds for A β positivity have been reported in detail elsewhere [24,36]. To estimate the proportion of

variance attributable to $A\beta_{42}$ -related longitudinal decline, we regressed out covariate effects, and then modeled the residuals as a function of $A\beta$ and $A\beta \times \text{age}$. Next, we plotted the generalized R^2 for these models (R^2_{GLMM}) [37]. To examine absolute effect sizes across outcomes, we plotted 95% CIs of the $\hat{\beta}_{A\beta \times \text{age}}$ terms obtained from linear mixed models of each outcome. Secondary validity analyses explored Spearman correlations between continuous $A\beta$ biomarker values and individual age-slope estimates for each outcome.

2.4.5. Sensitivity analyses

To examine the robustness of the IISD findings, we estimated mean IISD in a larger sample including 52 individuals that had previously been excluded due to a diagnosis of MCI or dementia during the study. We compared the average IISD for each outcome in this sample to the main findings and evaluated the differences in mean IISD between impaired and unimpaired individuals. In this expanded sample, we also compared IISD of all outcomes for a variety of risk groups to that observed in a lower-risk comparison group, as others have reported fluctuations in cognitive status in similar risk groups [4]. Parallel sensitivity analyses examined the robustness of the criterion validity findings to the removal of those with clinical impairment.

3. Results

3.1. Participants

Demographic information for the whole sample, the subset with CSF or PiB amyloid data, and the set who did not meet inclusion criteria are summarized in Table 1.

3.2. Convergent and discriminant validity

Intercorrelations among raw and composite scores are illustrated in Fig. 1. In general, scores related to executive function (STROOP, TMT-A, TMT-B, DIGSYM, THEO-EXEC-FN, EMP-EXEC-FN) were only weakly related to those in the episodic memory domains (AVLT-T, AVLT-D, LM-I, LM-II, BVMT-T, BVMT-D, THEO-IMM-LRN, THEO-DEL-REC, EMP-IMM-LRN, EMP-DEL-REC; median = 0.27, range = 0.07–0.41). Intercorrelations between memory-domain scores were stronger (median = 0.51, range = 0.27–0.97). Two raw scores in particular, MMSE and CFL, exhibited low correlations with all outcomes other than the related global composites (median = 0.24, range excluding related composites = 0.14–0.36). Intercorrelations were quite high among global composites (PACC4-MMSE, PACC4-CFL, PACC4-TMTB, PACC3, PACC3-TMTB; median = 0.9, range = 0.82–0.94) and between global and domain-specific composites (THEO-IMM-LRN, THEO-DEL-REC, THEO-EXEC-FN; median = 0.75, range = 0.62–0.86).

The matrix in Table 3 illustrates reliability and discriminant validity measures for three cognitive domains. Intra-class measures of reliability (within-domain, within-test) were reasonably high. However, the pattern of intercorrelations suggests a strong methods effect and relatively weak discriminant validity for the two memory domains. For executive function, within-domain, between-test correlations were similarly low, in line with other reports of high dispersion among executive function measures [38].

3.3. Intraindividual longitudinal variability

Fig. 2A illustrates intraindividual variability in each score over time, using the standardization sample of cognitively unimpaired individuals ($N = 1063$). Within domains, composites had lower IISDs than individual test raw scores. However, executive function raw and composite scores were less variable than scores from other domains, and some global composites as well. The MMSE raw score exhibited the largest IISD.

3.4. Criterion validity

Age-related slope estimates (Fig. 2B) for all outcomes were negative, indicating general decline with age. The two executive function composites (EMP-EXEC-FN and THEO-EXEC-FN), the DIGSYM raw score showed the most age-related change; slightly less was observed for the four global composites. The remaining composites and raw scores had slopes closer to zero.

The biomarker subsample ($N = 226$) showed a very similar IISD pattern (Fig. 3A). Fig. 3B–C illustrates two quantities related to criterion validity of each score. In few cases did the proportion of variance (generalized R^2) attributable to $A\beta$ positivity and its interaction with age exceed 0.02, indicating weak relationships between $A\beta$ positivity, cognition, and cognitive change in this largely cognitively unimpaired sample (Fig. 3B). Parameter estimates for the $A\beta$ positivity $\times$ age interaction (Fig. 3C) generally indicated worse age-related change in the $A\beta$ -positive group, but group differences were modest, with most confidence intervals including zero. Confidence intervals were smallest for executive-function measures and larger for other raw scores and empirical composites. All theoretical composites had point estimates on the larger end, and most global composites performed similarly.

In exploratory analyses (Supplementary Fig. 1), Spearman correlations between individual random slope estimates and three continuous $A\beta$ biomarkers were also most consistently visible for executive function measures ($\rho_{CSF-A\beta_{42}, DIGSYM} = 0.26$; $\rho_{CSF-A\beta_{42}/A\beta_{40}, DIGSYM} = 0.24$; $\rho_{CSF-A\beta_{42}/A\beta_{40}, EMP-EXEC-FN} = 0.21$; $\rho_{PiB-DVR, STROOP} = -0.23$). The remaining correlations were smaller ($-0.2 < \rho < 0.2$) and/or confined to a single biomarker ($\rho_{CSF-A\beta_{42}, EMP-LM} = 0.25$; $\rho_{CSF-A\beta_{42}, LM-I} = 0.22$), although all were in the expected direction, with better scores

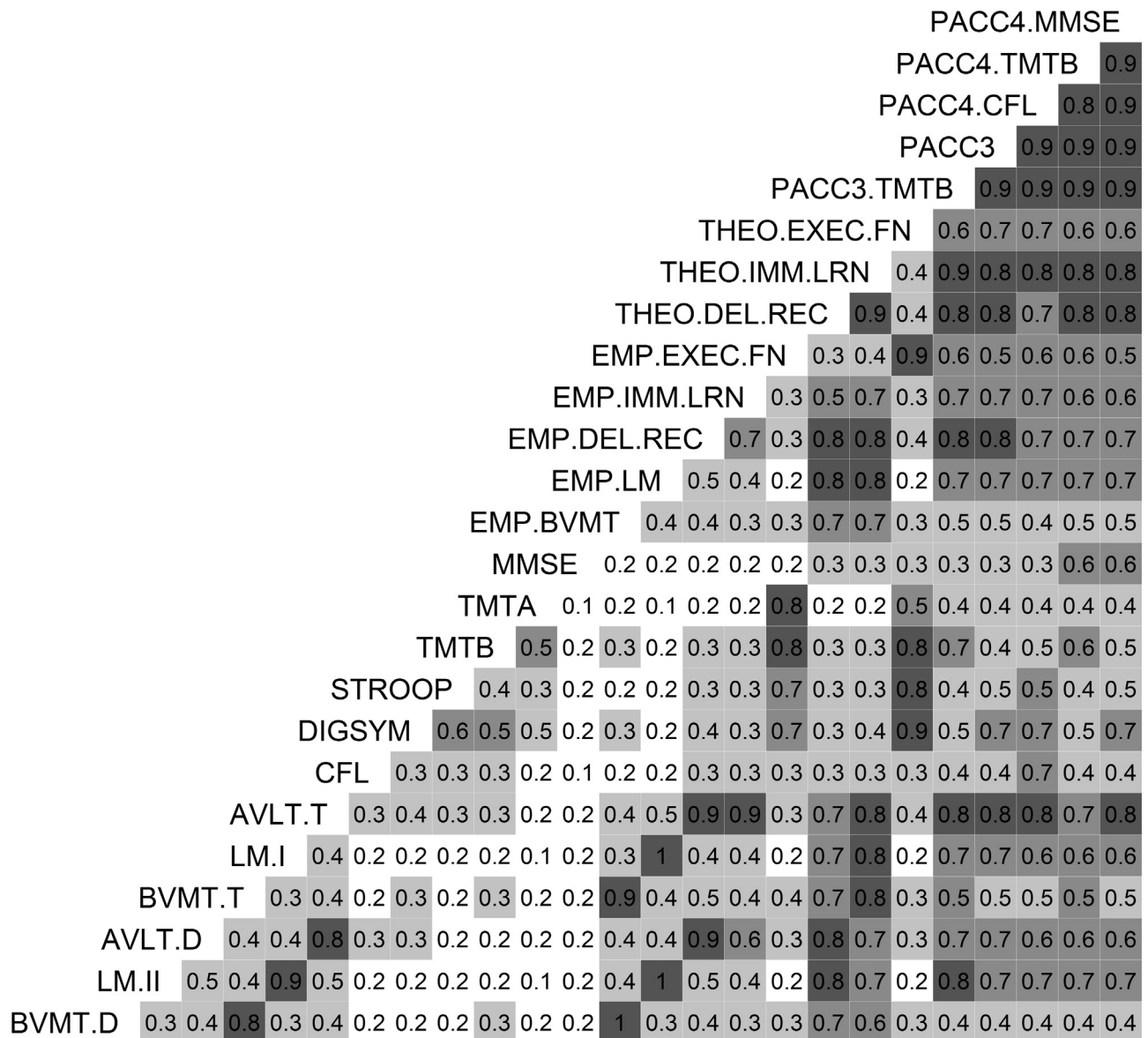


Fig. 1. Correlogram illustrating relationships between all outcomes. Darker shading indicates correlations closer to 1. Abbreviations: AVLT, Auditory-Verbal Learning Test; BVMT-R, Brief Visuospatial Memory Test–Revised; DEL-REC, delayed recall; EMP, empirical composites; EXEC-FN, executive function; IMM-LRN, immediate learning; LM, Logical Memory; MMSE, Mini-Mental State Exam; PACC, preclinical Alzheimer's cognitive composite; THEO, theoretical composites; TMT, Trail Making Test; DIGSYM, Digit Symbol subtest of the Wechsler Memory Scale–Revised.

associated with larger CSF-A β biomarker values and smaller [^{11}C]PiB distribution volume ratio values.

3.5. Sensitivity analyses

We recalculated IISD on a larger data set including participants with at least one diagnosis of clinical MCI or worse at any point during the study ($N = 1115$). Mean IISDs in this subsample were very similar to the standardization sample ($r = 0.997$), indicating low sensitivity of our results to this exclusion criterion. However, IISD values tended to be higher for the added participants, with greater discrepancies

for some outcomes (e.g., TMT-B, $\Delta_{IISD} = 0.57$) than others (STROOP, $\Delta_{IISD} \approx 0$). [Supplementary Fig. 2](#) illustrates the relationships between mean IISD in this sample and the group difference in IISDs between cognitively unimpaired participants and those with clinically significant cognitive impairment. The global composites tend to cluster in the quadrant with lower mean IISD and greater discrepancies between the clinical and nonclinical samples.

[Supplementary Fig. 3](#) illustrates IISD for each outcome in a healthy subgroup ($APOE\epsilon 3/\epsilon 3$ participants who were in good health at last visit and reported no clinical or psychiatric diagnosis at any point; [Supplementary Fig. 3](#), top row)

Table 3

Multitrait, multimethod matrix [32] evaluating the convergent and discriminant validity of the constructs represented by the immediate learning, delayed recall, and executive function theoretically derived composites

Raw scores	AVLT-T	AVLT-D	LM-I	LM-II	BVMT-T	BVMT-D	TMT-B	STROOP	DIGSYM
AVLT-T	0.68	0.42	<i>0.14</i>	0.15	<i>0.12</i>	0.07	0.03	0.05	0.06
AVLT-D		0.68	0.15	<i>0.17</i>	0.12	<i>0.05</i>	0.03	0.06	0.01
LM-I			0.63	0.74	<i>0.13</i>	0.08	0.04	0.07	0.09
LM-II				0.68	0.16	<i>0.11</i>	0.05	0.05	0.06
BVMT-T					0.59	0.61	0.08	–0.02	0.02
BVMT-D						0.55	0.06	0.01	0.03
TMT-B							0.64	<i>0.06</i>	<i>0.11</i>
STROOP								0.82	<i>0.22</i>
DIGSYM									0.84

Abbreviations: AVLT, Auditory-Verbal Learning Test; BVMT, Brief Visuospatial Memory Test–Revised; LM, Logical Memory; TMT, Trail Making Test; STROOP, Stroop test, Color–Word Interference; DIGSYM, Digit Symbol subtest of the Wechsler Adult Intelligence Scale–Revised.

NOTE. Main diagonal represents intraclass correlation (ICC) for within-subject variability. Off-diagonal represents repeated measures correlations between tests, adjusting for subject-level variance. Cells denoting pairwise comparisons within a test are bolded; cells denoting comparisons within a domain are italicized.

and several risk groups (*APOE*ε4 carriers; those reporting a major psychiatric diagnosis; those reporting fair or poor health at last visit; and those receiving a clinical consensus diagnosis at any time). In our sample, those with clinical MCI or worse appeared to have slightly elevated IISD on some outcomes. In contrast to Sugarman [4], other subgroups showed variability similar to the healthy subgroup.

Sensitivity analyses for our criterion validity findings, in which clinically impaired individuals were removed from the biomarker subset, also showed little difference from the primary analyses, with high correlations between two estimates of IISD (0.997), generalized R^2 (0.988), and $\beta_{A\beta \times \text{age}}$ (0.984).

4. Discussion

In a sample of over 1000 cognitively unimpaired late middle-aged adults, we observed that global and theoretically derived domain-specific composites generally exhibited lower variability and stronger relationships with age and $A\beta$ compared with raw scores or to empirically derived composites [11,29]. This is broadly consonant with other findings [8,10]. Although the global composites excluding MMSE exhibited slightly smaller IISDs (Fig. 2A, 3A) and stronger relationships with $A\beta$ (Fig. 3B, C), these differences might not replicate in other samples. The key feature distinguishing global and theoretical composites from other scores is that these composites average across tests which load on distinct factors [11,29]. Variability induced by poor performance on only one test from a given theoretical domain is reduced, allowing time trends to become more visible.

Others have reported associations between intraindividual variability and cognitive impairment [1–3] or other neuropsychiatric problems [4]. We therefore conducted primary analyses in a sample without clinically significant

cognitive impairment to simplify the interpretation of variability. In follow-up analyses, we wondered whether those measures with low mean IISD values in a healthy sample would be sensitive enough to early change in those who are impaired. Indeed, in a sensitivity analysis on an expanded sample, mean IISD values for each outcome were quite similar, and some lower-IISD measures nevertheless evinced higher intraindividual variability in a subsample receiving a clinical diagnosis of MCI or worse during follow-up. However, no evidence of greater cognitive variability in other risk groups was observed.

The discriminant validity evidence for separate immediate learning and delayed recall factors in this data set is quite weak (Table 3). This was moderately surprising, as previous analyses in this sample suggested separate immediate and delayed memory components for the AVLT [29]. A reanalysis incorporating single-trial-level data for each memory test might more closely mirror the earlier result. However, given the high correlation observed between the two theoretical memory composites (Fig. 1), it may be worth considering a memory composite incorporating both immediate and delayed information.

The strong correlations among global composites are of practical importance for researchers wishing to compare results across studies, as variation across neuropsychological testing batteries is a common feature. These results confirm and extend the work of Donohue and colleagues to create a composite that can be used with modification in multiple cohorts [13]. The scientific community has recently acknowledged the importance of replication studies in neuropsychology [7]; thus, having a class of lower-inconsistency, high-criterion-validity composites that can be modified based on availability of inputs is beneficial.

The superiority of executive function measures on both consistency and some criterion validity measures was unexpected, as changes in memory are often thought to be the

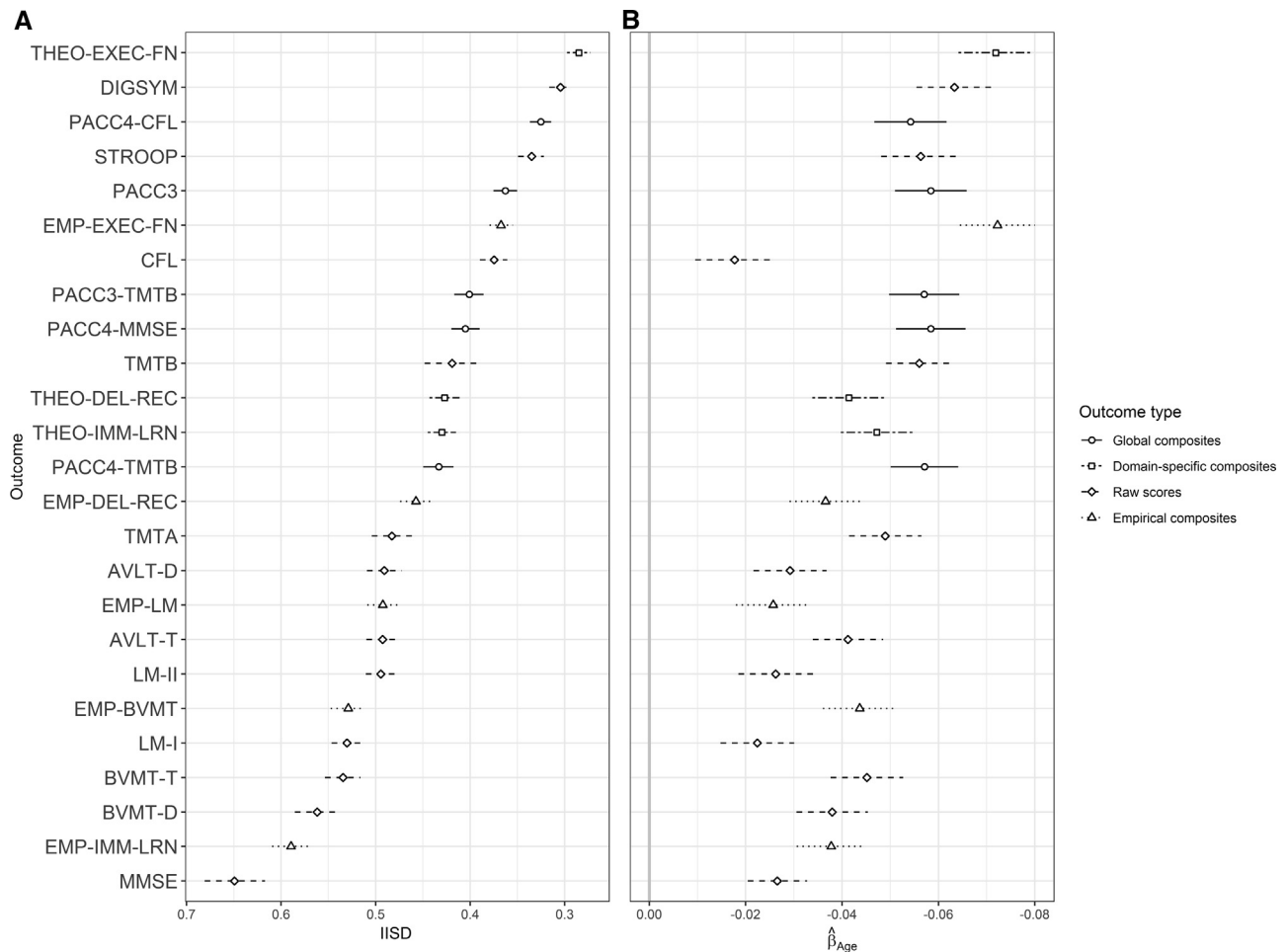


Fig. 2. Performance of individual cognitive scores on two metrics of interest in entire sample ($N = 1063$). The y-axis is ordered by ascending mean IISD. Each x-axis has been oriented such that scores further to the right indicate more favorable measurement characteristics (A: lower IISD; B: greater sensitivity to age-related decline). (A) Mean intraindividual standard deviation (IISD) for all outcomes, with bootstrapped 95% confidence intervals. (B) Parameter estimate describing age-related change from full models of cognitive outcomes including other covariates (sex, education, baseline literacy, and prior practice with the battery). Error bars represent parametric 95% confidence intervals around the estimate. Abbreviations: AVLT, Auditory-Verbal Learning Test; BVMT-R, Brief Visuospatial Memory Test–Revised; DEL-REC, delayed recall; EMP, empirical composites; EXEC-FN, executive function; IMM-LRN, immediate learning; LM, Logical Memory; MMSE, Mini-Mental State Exam; PACC, preclinical Alzheimer's cognitive composite; THEO, theoretical composites; TMT, Trail Making Test; DIGSYM, Digit Symbol subtest of the Wechsler Memory Scale–Revised.

earliest cognitive signal associated with AD [9]. Some other reports suggest executive function changes in early AD [39,40], and the relationship between lower executive function and biomarkers of brain amyloidosis has been observed before in this preclinical cohort [12]. However, we caution that some of what appears in this article to be a consistency advantage of executive function tests may be principally a function of normal aging [41], rather than disease-related processes, as outcomes that change more reliably with age will look superior by our inconsistency metric. The slight apparent advantage of executive function scores in relating to biomarkers (Fig. 3B; Supplementary Fig. 1) was not consistent across all metrics (Fig. 3C) [12] and should not be overinterpreted, except as evidence that such measures are appropriate to include in a comprehensive

cognitive battery. We will re-examine this question directly once more of the WRAP cohort has reached a clinical endpoint.

4.1. Limitations

In these analyses, we did not perform formal hypothesis tests comparing composites to each other, and the confidence intervals we present (e.g., around beta estimates) have not been adjusted for multiple comparisons. We chose this approach because in a clinical trial setting, one or two outcomes would be selected as primary, so what researchers most need is not the proof that these outcomes are statistically distinguishable—they may not be—but instead, an understanding of the range of longitudinal

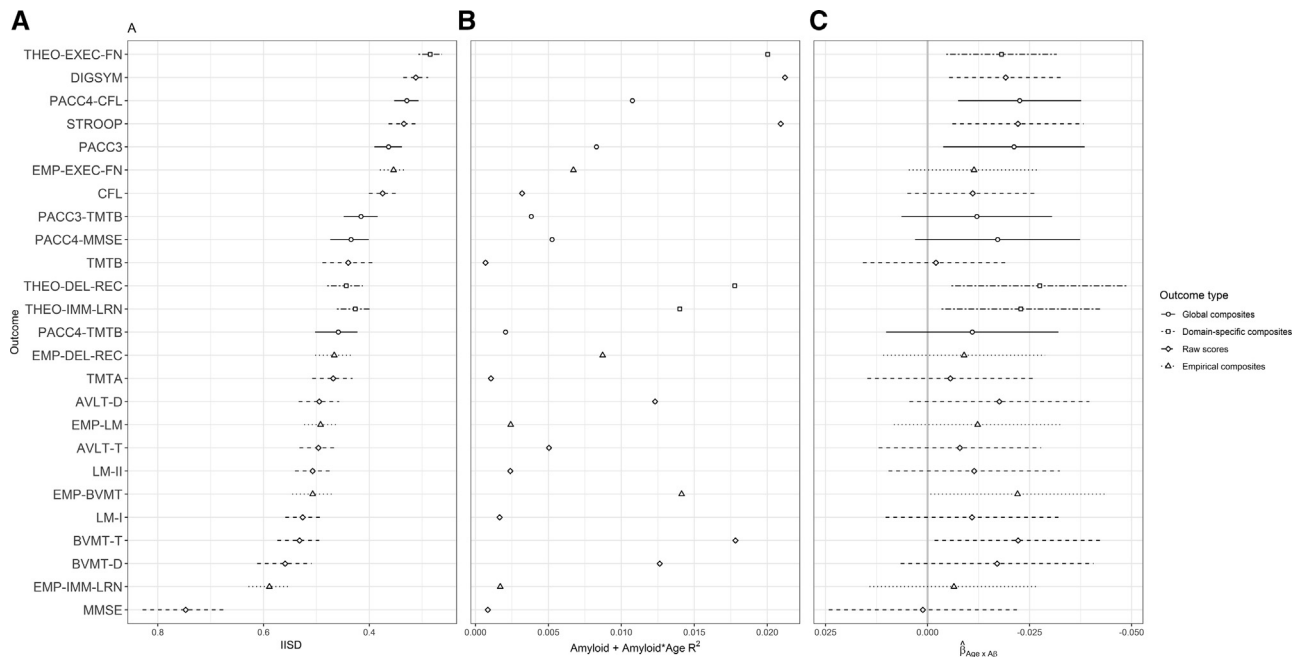


Fig. 3. Performance of individual cognitive scores on three metrics of interest in the subsample having biomarkers ($N = 226$). The y-axis preserves the order of Fig. 2A. Each x-axis has been oriented such that scores further to the right indicate more favorable measurement characteristics (A: lower IISD; B-C: greater sensitivity to age-related decline). (A) Mean intraindividual standard deviation (IISD) for all outcomes, with bootstrapped 95% confidence intervals. (B) The proportion of variance (R^2_{GLMM}) [37] in cognitive outcomes attributable to $A\beta$ and its interaction with age, after adjusting for standard covariates (age, sex, education, baseline literacy, and prior practice with the battery). (C) Parameter estimate describing age $\times$ $A\beta$ interaction from full models of cognitive outcomes including covariates and $A\beta$. Larger negative values for this parameter estimate suggest worse age-related change in $A\beta$ -positive individuals. Error bars represent parametric 95% confidence intervals around the estimate. Abbreviations: AVLT, Auditory-Verbal Learning Test; BVMT-R, Brief Visuospatial Memory Test-Revised; DEL-REC, delayed recall; EMP, empirical composites; EXEC-FN, executive function; IMM-LRN, immediate learning; LM, Logical Memory; MMSE, Mini-Mental State Exam; PACC, preclinical Alzheimer's cognitive composite; THEO, theoretical composites; TMT, Trail Making Test; DIGSYM, Digit Symbol subtest of the Wechsler Memory Scale-Revised.

variation and strength of relationship with criterion variables that they might expect for each, in samples similar to WRAP.

The tests covered by our analyses also did not span the entire range of cognitive function. In particular, confrontation naming, assessed in WRAP using the Boston Naming Test [42], was not considered. Previous analyses in this cohort have suggested there is not yet enough variability in this measure for it to be a meaningful differentiator [43]. Instead, we focused on measures that were components of one of several composites of interest to us, so that we could more easily make relevant comparisons.

4.2. Conclusion and future directions

These results reinforce the need for careful selection of cognitive outcomes when designing studies, and provide support for composite over raw scores because of lower longitudinal intraindividual variability and stronger relationships with AD biomarkers. Future work building on these

findings will examine the relevance of this inconsistency measure to clinical trial planning.

Acknowledgments

This research was supported by the National Institutes of Health awards R01 AG027161, R01 AG021155, R01 AG054047, R01 AG037639, R01 AG054059, P50 AG033514, and UL1 TR000427, and by donor funds including the Wisconsin Alzheimer's Institute Lou Holland Fund and contributions from anonymous donors. Portions of this research were supported by resources at the Wisconsin Alzheimer's Institute, the Wisconsin Alzheimer's Disease Research Center and the Geriatric Research Education and Clinical Center of the William S. Middleton Memorial Veterans Hospital, Madison, WI. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NIH or the Veterans Administration. The authors gratefully acknowledge the WRAP study team members who have

carefully acquired the longitudinal data, and the WRAP participants who make this research possible.

Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.dadm.2018.11.006>.

RESEARCH IN CONTEXT

1. Systematic review: We used PubMed to find articles discussing intraindividual variability and the construction of composite scores. Interest in composites in particular is growing and several key articles are cited, with special emphasis on the work by Donohue et al. describing the Preclinical Alzheimer's Cognitive Composite.
2. Interpretation: We used the longitudinal intraindividual standard deviation to quantify the variability of different scores in the same set of participants. Like other research groups using different metrics, we found composites to be advantageous.
3. Future directions: Assessing criterion validity in a middle-aged cohort is difficult because of the lack of true clinical endpoints. Future work should examine whether low-IISD measures like the selected composites are also good prognostic indicators of the eventual development of dementia.

References

- [1] Burton CL, Strauss E, Hultsch DF, Moll A, Hunter MA. Intraindividual variability as a marker of neurological dysfunction: A comparison of Alzheimer's disease and Parkinson's disease. *J Clin Exp Neuropsychol* 2006;28:67–83.
- [2] Salthouse TA. Psychometric properties of within-person across-session variability in accuracy of cognitive performance. *Assessment* 2012;19:494–501.
- [3] Hultsch DF, MacDonald SW, Hunter MA, Levy-Bencheton J, Strauss E. Intraindividual variability in cognitive performance in older adults: Comparison of adults with mild dementia, adults with arthritis, and healthy adults. *Neuropsychology* 2000;14:588–98.
- [4] Sugarman MA, Alosco ML, Tripodis Y, Steinberg EG, Stern RA. Neuropsychiatric Symptoms and the Diagnostic Stability of Mild Cognitive Impairment. *J Alzheimers Dis* 2018;62:1841–55.
- [5] Johnson SC, Kosciak RL, Jonaitis EM, Clark LR, Mueller KD, Berman SE, et al. The Wisconsin Registry for Alzheimer's Prevention: A review of findings and current directions. *Alzheimers Dement* 2018;10:130–42.
- [6] Weintraub S, Besser L, Dodge HH, Teylan M, Ferris S, Goldstein FC, et al. Version 3 of the Alzheimer Disease Centers' Neuropsychological Test Battery in the Uniform Data Set (UDS). *Alzheimer Dis Assoc* 2018;32:10–7.
- [7] Gelman A, Geurts HM. The statistical crisis in science: How is it relevant to clinical neuropsychology? *Clin Neuropsychol* 2017;31:1000–14.
- [8] Donohue MC, Sperling RA, Salmon DP, Rentz DM, Raman R, Thomas RG, et al. The preclinical Alzheimer cognitive composite: Measuring amyloid-related decline. *JAMA Neurol* 2014;71:961–70.
- [9] Mortamais M, Ash JA, Harrison J, Kaye J, Kramer J, Randolph C, et al. Detecting cognitive changes in preclinical Alzheimer's disease: A review of its feasibility. *Alzheimers Dement* 2017;13:468–92.
- [10] Kozauer N, Katz R. Regulatory innovation and drug development for early-stage Alzheimer's disease. *N Engl J Med* 2013;368:1169–71.
- [11] Dowling NM, Hermann B, La Rue A, Sager MA. Latent structure and factorial invariance of a neuropsychological test battery for the study of preclinical Alzheimer's disease. *Neuropsychology* 2010;24:742–56.
- [12] Clark LR, Racine AM, Kosciak RL, Okonkwo OC, Engelman CD, Carlsson CM, et al. Beta-amyloid and cognitive decline in late middle age: Findings from the Wisconsin Registry for Alzheimer's Prevention study. *Alzheimers Dement* 2016;12:805–14.
- [13] Donohue MC, Sun C-K, Raman R, Insel PS, Aisen PS, AN-ADNI, et al. Cross-validation of optimized composites for preclinical Alzheimer's disease. *Alzheimers Dement* 2017;3:123–9.
- [14] Kane M, Case SM. The reliability and validity of weighted composite scores. *Appl Meas Educ* 2004;17:221–40.
- [15] Schmidt M. Rey Auditory Verbal Learning Test: A handbook. Los Angeles, CA: Western Psychological Services; 1996.
- [16] Wechsler D. Wechsler Memory Scale - Revised. New York, NY: Harcourt Brace Jovanovich, Inc. for the Psychological Corporation; 1987.
- [17] Benedict R. Brief Visuospatial Memory Test - Revised. Odessa, FL: Psychological Assessment Resources, Inc.; 1997.
- [18] Trenerry B, Crosson. The Stroop Neuropsychological Screening Test. Odessa, FL: Psychological Assessment Resources, Inc.; 1989.
- [19] Reitan RM. Validity of the Trail Making Test as an indicator of organic brain damage. *Percept Mot Skills* 1958;8:271–6.
- [20] Wechsler D. Wechsler Adult Intelligence Scale - III. San Antonio, TX: The Psychological Corporation; 1997.
- [21] Rosen WG. Verbal fluency in aging and dementia. *J Clin Exp Neuropsychol* 1980;2:135–46.
- [22] Folstein MF, Folstein SE, McHugh PR. "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res* 1975;12:189–98.
- [23] Wilkinson GS, ed. The Wide Range Achievement Test: Manual. Third ed. Wilmington, DE: Jastak Association; 1993.
- [24] Clark L, Berman S, Norton D, Kosciak R, Jonaitis E, Blennow K, et al. Age-accelerated cognitive decline in asymptomatic adults with CSF β -amyloid. *Neurology* 2018;90:e1306–15.
- [25] Johnson SC, Christian BT, Okonkwo OC, Oh JM, Harding S, Xu G, et al. Amyloid burden and neural function in people at risk for Alzheimer's Disease. *Neurobiol Aging* 2014;35:576–84.
- [26] Logan J, Fowler JS, Volkow ND, Wang GJ, Ding YS, Alexoff DL. Distribution volume ratios without blood sampling from graphical analysis of PET data. *J Cereb Blood Flow* 1996;16:834–40.
- [27] Lopresti BJ, Klunk WE, Mathis CA, Hoge JA, Ziolkowski SK, Lu X, et al. Simplified quantification of Pittsburgh Compound B amyloid imaging PET studies: A comparative analysis. *J Nucl Med* 2005;46:1959–72.
- [28] Sprecher KE, Bendlin BB, Racine AM, Okonkwo OC, Christian BT, Kosciak RL, et al. Amyloid burden is associated with self-reported sleep in nondemented late middle-aged adults. *Neurobiol Aging* 2015;36:2568–76.
- [29] Kosciak RL, La Rue A, Jonaitis EM, Okonkwo OC, Johnson SC, Bendlin BB, et al. Emergence of mild cognitive impairment in late middle-aged adults in the Wisconsin Registry for Alzheimer's Prevention. *Demen Geriatr Cogn Disord* 2014;38:16–30.
- [30] Mormino EC, Papp KV, Rentz DM, Donohue MC, Amariglio R, Quiroz YT, et al. Early and late change on the preclinical Alzheimer's cognitive composite in clinically normal older individuals with elevated amyloid- β . *Alzheimers Dement* 2017;13:1004–12.
- [31] Lim YY, Snyder PJ, Pietrzak RH, Ukiqi A, Villemagne VL, Ames D, et al. Sensitivity of composite scores to amyloid burden in preclinical

- Alzheimer's disease: Introducing the Z-scores of Attention, Verbal fluency, and Episodic memory for Nondemented older adults composite score. *Alzheimers Dement* 2015;2:19–26.
- [32] Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull* 1959;56:81–105.
- [33] Bakdash JZ, Marusich LR. Repeated Measures Correlation. *Front Psychol* 2017;8:456.
- [34] Rmcorr: Repeated Measures Correlation [computer program]; 2018. Bakdash JZ and Marusich LR. Available at: <https://cran.r-project.org/web/packages/rmcorr/index.html>
- [35] Golay P, Fagot D, Lecerf T. Against coefficient of variation for estimation of intraindividual variability with accuracy measures. *Tutor Quant Methods Psychol* 2013;9:6–14.
- [36] Racine AM, Clark LR, Berman SE, Kosciak RL, Mueller KD, Norton D, et al. Associations between performance on an abbreviated cogstate battery, other measures of cognitive function, and biomarkers in people at risk for Alzheimer's disease. *J Alzheimers Dis* 2016;54:1395–408.
- [37] Nakagawa S, Schielzeth H. A general and simple method for obtaining R² from generalized linear mixed-effects models. *Methods Ecol Evol* 2013;4:133–42.
- [38] Buczyłowska D, Petermann F. Intraindividual variability in executive function performance in healthy adults: cross-sectional analysis of the nab executive functions module. *Front Psychol* 2018; 9:329.
- [39] Lafleche G, Albert M. Executive function deficits in mild Alzheimers-Disease. *Neuropsychology* 1995;9:313–20.
- [40] Baudic S, Barba GD, Thibaudet MC, Smagghe A, Remy P, Traykov L. Executive function deficits in early Alzheimer's disease and their relations with episodic memory. *Arch Clin Neuropsychol* 2006;21:15–21.
- [41] Anderson ND, Craik FIM. 50 Years of Cognitive Aging Theory. *J Gerontol Ser B Psychol Sci Soc Sci* 2017;72:1–6.
- [42] Kaplan H, Goodglass. Boston Naming Test. Philadelphia, PA: Lea & Febiger; 1983.
- [43] Kosciak RL, Norton DL, Allison SL, Jonaitis EM, Clark LR, Mueller KD, et al. Characterizing the effects of sex, APOE ε4, and literacy on mid-life cognitive trajectories: Application of Information-Theoretic model averaging and multi-model inference techniques to the Wisconsin Registry for Alzheimer's Prevention Study n.d.