Neuropsychology

Harmonizing the Preclinical Alzheimer Cognitive Composite for Multicohort Studies

Olivia L. Hampton, Shubhabrata Mukherjee, Michael J. Properzi, Aaron P. Schultz, Paul K. Crane, Laura E. Gibbons, Timothy J. Hohman, Paul Maruff, Yen Ying Lim, Rebecca E. Amariglio, Kathryn V. Papp, Keith A. Johnson, Dorene M. Rentz, Reisa A. Sperling, and Rachel F. Buckley

Online First Publication, July 21, 2022. http://dx.doi.org/10.1037/neu0000833

AMERICAN PSYCHOLOGICAL ASSOCIATION

Neuropsychology

ISSN: 0894-4105

https://doi.org/10.1037/neu0000833

Harmonizing the Preclinical Alzheimer Cognitive Composite for Multicohort Studies

Olivia L. Hampton[1], Shubhabrata Mukherjee[2], Michael J. Properzi[1], Aaron P. Schultz[1], Paul K. Crane[2], Laura E. Gibbons[2], Timothy J. Hohman[3], Paul Maruff[4], Yen Ying Lim[6], Rebecca E. Amariglio[1, 5], Kathryn V. Papp[1, 5], Keith A. Johnson[5, 7], Dorene M. Rentz[1, 5], Reisa A. Sperling[1, 5], and Rachel F. Buckley[1, 5, 8]

[1] Department of Neurology, Harvard Medical School, Massachusetts General Hospital, Boston, Massachusetts, United States
[2] Department of Medicine, Division of General Internal Medicine, University of Washington
[3] Department of Neurology, Vanderbilt University Medical Center, Nashville, Tennessee, United States
[4] Cogstate Ltd., Melbourne, Victoria, Australia
[5] Department of Neurology, Brigham and Women's Hospital, Center for Alzheimer Research and Treatment, Boston, Massachusetts, United States
[6] Turner Institute for Brain and Mental Health, Monash University, Melbourne, Victoria, Australia
[7] Department of Radiology, Harvard Medical School, Massachusetts General Hospital, Boston, Massachusetts, United States
[8] Melbourne School of Psychological Science, University of Melbourne

on behalf of the A4 Study, the Alzheimer's Disease Neuroimaging Initiative, the Australian Imaging Biomarkers and Lifestyle Study of Ageing, and the Harvard Aging Brain Study

Objectives: Studies are increasingly examining research questions across multiple cohorts using data from the preclinical Alzheimer cognitive composite (PACC). Our objective was to use modern psychometric approaches to develop a harmonized PACC. ***Method:*** We used longitudinal data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), Harvard Aging Brain Study (HABS), and Australian Imaging, Biomarker and Lifestyle Study of Ageing (AIBL) cohorts ($n = 2{,}712$). We further demonstrated our method with the Anti-Amyloid Treatment of Asymptomatic Alzheimer's Disease (A4) Study prerandomized data ($n = 4{,}492$). For the harmonization method, we used confirmatory factor analysis (CFA) on the final visit of the longitudinal cohorts to determine parameters to generate latent PACC (lPACC) scores. Overlapping tests across studies were set as "anchors" that tied cohorts together, while parameters from unique tests were freely estimated. We performed validation analyses to assess the performance of lPACC versus the common standardized PACC (zPACC). ***Results:*** Baseline (BL) scores for the zPACC were centered on zero, by definition. The harmonized lPACC did not define a common mean of zero and demonstrated differences in baseline ability levels across the cohorts. Baseline lPACC slightly outperformed zPACC in the prediction of progression to dementia. Longitudinal change in the lPACC was more constrained and less variable relative to the zPACC. In combined-cohort analyses, longitudinal lPACC slightly outperformed longitudinal zPACC in its association with baseline β-amyloid status. ***Conclusions:*** This study proposes procedures for harmonizing the PACC that make fewer strong assumptions than the zPACC, facilitating robust

Rachel F. Buckley https://orcid.org/0000-0002-5356-5537

The authors thank the participants who volunteered their valuable time to these studies. The Harvard Aging Brain Study is funded by the National Institute on Aging (P01AG036694) with additional support from several philanthropic organizations. Data collection and sharing for this project were funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI; National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute (ARTI) at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. Data used in the preparation of this article were obtained from the Australian Imaging, Biomarker and Lifestyle Study of Ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) which was made available at the ADNI database (www.loni.usc.edu/ADNI). The A4 study is a secondary prevention trial in preclinical Alzheimer's disease, aiming to slow cognitive decline associated with brain amyloid accumulation in clinically normal older individuals. The A4 study is funded by a public–private–philanthropic partnership, including funding from the National Institutes of Health–National Institute on Aging, Eli Lilly and Company, Alzheimer's

continued

multicohort analyses. This implementation of item response theory lends itself to adapting across future cohorts with similar composites.

Key Points

Question: Does a harmonized version of the preclinical Alzheimer cognitive composite (PACC), computed using item response theory, perform better than a traditional standardized version? ***Findings:*** The harmonized PACC reveals inherent baseline differences between the cohorts that the standardized PACC masks. The harmonized PACC performs similarly, albeit slightly better, as a longitudinal outcome variable relative to the standardized PACC. ***Importance:*** This harmonized PACC can be used in multicohort analyses applies modern psychometric approaches to translating this composite across cohorts. ***Next Steps:*** Our next steps will be to gather more diverse cohorts in order to build a more generalizable sample for the legacy model.

Keywords: cognition, Alzheimer's disease, harmonization

The preclinical Alzheimer cognitive composite (PACC) is derived from a small set of neuropsychological tests that measure episodic memory, executive function, and general mental status across observational studies and secondary clinical trials of preclinical Alzheimer's disease (AD; Donohue et al., 2014; Donohue, Sun, et al., 2017). Decline on the PACC reflects the early changes that characterize preclinical AD (Donohue et al., 2014; Papp et al., 2017). The original composite combines measures of episodic memory, executive function, and global cognition, with a heavier weighting toward episodic memory (Donohue et al., 2014). More recent iterations of the PACC include a measure of verbal fluency (PACC-5), which has increased the explained variance associated with Aβ-related cognitive decline (Lim et al., 2016; Papp et al., 2017).

Studies that have created a PACC include study-specific tests for each cognitive domain included in the PACC; the only exception to this rule is the Mini-Mental State Examination (MMSE; Folstein et al., 1975), which is included in the neuropsychological battery of most cohorts. Historically, calculating the PACC for an individual required standardization of performance on each neuropsychological test at each assessment, to that of the study sample at baseline (BL). Standardized scores for each test are then averaged (Mormino et al., 2017; Papp et al., 2020) or summed (Donohue et al., 2014) for each assessment. Although the cognitive test items used to form the PACC are well validated, sum scoring or averaging these items makes important assumptions about the psychometric properties of the tests which may then render the composite less informative or sensitive to AD-related processes. This limits the direct comparison of scores between studies and limits generalizability to other samples (McNeish & Wolf, 2020). While studies exist that have used the standardized version of the PACC across cohorts (Buckley et al., 2018; Insel et al., 2019; Mormino et al., 2017; Papp et al., 2020), it behooves the field to investigate more sophisticated

Association, Accelerating Medicines Partnership, GHR Foundation, an anonymous foundation, and additional private donors, with in-kind support from Avid and Cogstate. The companion observational Longitudinal Evaluation of Amyloid Risk and Neurodegeneration (LEARN) Study is funded by the Alzheimer's Association and GHR Foundation. The A4 and LEARN Studies are led by Reisa A. Sperling at Brigham and Women's Hospital, Harvard Medical School and Paul Aisen at the ATRI, University of Southern California. The A4 and LEARN studies are coordinated by ATRI at the University of Southern California, and the data are made available through the Laboratory for Neuro Imaging at the University of Southern California. The participants screening for the A4 study provided permission to share their de-identified data in order to advance the quest to find a successful treatment for Alzheimer's disease. The authors would like to acknowledge the dedication of all the participants, the site personnel, and all of the partnership team members who continue to make the A4 and LEARN Studies possible. The complete A4 study team list is available at a4study.org/a4-study-team. Rachel F. Buckley is supported by a K99/R00 award from NIA (R00AG061238-03), an Alzheimer's Association Research Fellowship (AARF-20-675646), and philanthropic support.

Olivia L. Hampton played a lead role in visualization, writing of original draft, and writing of review and editing, a supporting role in methodology and validation, and an equal role in data curation and formal analysis. Shubhabrata Mukherjee played a supporting role in conceptualization, formal analysis, visualization, and writing of review and editing, and an equal role in methodology. Michael J. Properzi played a supporting role in methodology and writing of review and editing. Aaron P. Schultz played a supporting role in methodology and writing of review and editing. Paul K. Crane played a supporting role in methodology and validation and an equal role in writing of review and editing. Laura E. Gibbons played a supporting role in writing of review and editing. Timothy J. Hohman played a supporting role in writing of review and editing. Paul Maruff played a supporting role in writing of review and editing. Yen Ying Lim played a supporting role in writing of review and editing. Rebecca E. Amariglio played a supporting role in writing of review and editing. Kathryn V. Papp played a supporting role in writing of review and editing. Keith A. Johnson played a supporting role in writing of review and editing. Dorene M. Rentz played a supporting role in writing of review and editing. Reisa A. Sperling played a supporting role in writing of review and editing. Rachel F. Buckley played a lead role in conceptualization, methodology, supervision, and writing of review and editing, a supporting role in visualization and writing of original draft, and an equal role in data curation, formal analysis, and validation.

A complete listing of the Anti-Amyloid Treatment of Asymptomatic Alzheimer's Disease (A4) Study investigators can be found at http://nmr.mgh.harvard.edu/lab/harvardagingbrain/aboutus, Harvard Aging Brain Study (HABS) investigators can be found at http://nmr.mgh.harvard.edu/lab/harvardagingbrain/aboutus, AIBL researchers are listed at www.aibl.csiro.au. Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Correspondence concerning this article should be addressed to Rachel F. Buckley, Department of Neurology, Harvard Medical School, Massachusetts General Hospital, 149 13th Street, Charlestown, MA 02129. Email: rfbuckley@mgh.harvard.edu

cognitive harmonization approaches to determine if violation of the strong assumptions required for an average or sum of standardized scores to be an appropriate choice for obtaining a composite score for the PACC is associated with lower validity than other psychometric modeling choices that make fewer strong assumptions.

The strict assumptions required for standardization include equal weighting of tests, regardless of difficulty level, and equality in item-level scores of the same test across cohorts (Schneider & Goldberg, 2020). Such assumptions, when not met, may obscure the detection of true change (Schneider & Goldberg, 2020). Critically, standardization cannot guarantee that incremental change in one cohort reflects the same monotonic change across other cohorts when test versions or components differ across cohorts. This means that a one-unit change or difference in Study A does not have the same meaning as a one-unit change or difference in Study B, which is contrary to the goals of efforts to harmonize composite scoring for the studies employing the PACC. Further, although episodic memory and executive function domains are well represented in each PACC variant, the neuropsychological tests representing these domains may not exhibit measurement equivalence across cohorts. That is, different tests may be used to reflect the episodic memory construct, which themselves are not equivalent. This issue is highly similar to the measurement issue highlighted by Chapman and Chapman in the 1970s (Chapman & Chapman, 1978), where within a diagnostic battery, different measurement properties could lead to different conclusions about the diagnosis. Here, the issue is differences in measurement properties of subsets of the PACC that are different across different study cohorts, as opposed to the cross-domain differences highlighted by Chapman and Chapman. The general concerns are similar. The commonly used zPACC produces a veneer of equality as a composite score that appears to have the same units for each domain within each study and to have the same units across studies.

Alternative modeling approaches of standardization have been proposed to ensure cognitive data harmonization (Chan et al., 2015; Gibbons et al., 2011; Gross et al., 2015; Park et al., 2012). One approach is to treat the PACC as a latent trait that is anchored by tests that are shared across cohorts. Item response theory (IRT), outlined in Crane et al. (2008), asserts that anchor items facilitate better longitudinal modeling with data-driven weighting of test components and allow for the parameterization of such anchor items. This approach also enables direct comparability between cohorts.

One major advantage of using IRT models is the ability to use item-level (granular) data to estimate each of the item parameters to harmonize a latent trait across the cohorts. This sits in opposition to standardization approaches that use the total test score. The use of total (sum) scores again makes strong assumptions, here about the linearity of the scaling metric. When these assumptions have been tested with data they have been found to be violated (Crane et al., 2008). Furthermore, IRT models do not assume that all indicators have equal difficulty. For example, it has long been appreciated that different semantic categories in verbal fluency tasks may have different difficulty levels (Laiacona et al., 1998). IRT parameterization allows for different item parameters to account for those different difficulty levels; these differences can be masked by a total score (or equivalently by an average).

The simplest IRT models are single-factor models, where all the covariation among indicators of a latent trait is modeled as being due to relationships with the common factor. IRT models can also accommodate a secondary domain structure to model additional factors to explain covariance across items within tests. For example, in modeling memory, if we administer a multitrial word list learning and recall task along with other items tapping memory, we expect scores from each learning and recall trial to be more closely correlated with each other than with the other items tapping memory. This correlation structure would reflect the fact that the same list of words is assessed in each learning and recall trial, so we would expect correlation to be tighter across those trials than with any other indicator of memory. The secondary domain structures we employ model this additional source of covariation. For our purposes, this secondary factor is a nuisance. The secondary factor accounts for the fact that particular words on a word list may be more salient for some individuals than for others, leading to differential ability on the word list method factor. Our modeling goal was to capture the general factor that reflects overall cognition as captured by all of the items, accounting for methods effects with secondary domain structures as needed. Failing to include a word list-specific factor would overestimate the strength of association for each learning trial on the overall memory factor, as there is no other place in the model for the word list-specific correlation to go than on the general memory factor. Modeling the word list-specific factor produces scores generated from a model that comes closer to the theorized relationships among these items. A simple total or average across all of the learning trials makes much stronger assumptions, many of which are in conflict with current scientific theories about cognition (Borsboom, 2005).

IRT methods have been implemented to compute composite scores of memory and executive function in cohorts of AD research for integration across data sets (Dowling et al., 2010; Dumitrescu et al., 2020; Gross, Sherva, et al., 2014; Langa et al., 2020; Mukherjee et al., 2020). We seek to expand this technique to the PACC, as researchers increasingly examine modified versions of this composite within new cohorts (Betthauser et al., 2020; Buckley et al., 2018; Burnham et al., 2016; Chhetri et al., 2017; Jessen et al., 2018; Papp et al., 2020), and attempt to investigate PACC change across cohorts (Buckley et al., 2018; Papp et al., 2020). To build upon the development of the original PACC publications (Donohue et al., 2014), we have developed a harmonized PACC score across Alzheimer's Disease Neuroimaging Initiative (ADNI), Harvard Aging Brain Study (HABS), and Australian Imaging, Biomarker and Lifestyle Study of Ageing (AIBL) data. We also demonstrate the flexibility to incorporate other data sets as they become available, by calculating a harmonized PACC score for the baseline Anti-Amyloid Treatment of Asymptomatic Alzheimer's Disease (A4) Study data set. In this study, we detail the method and then compare the performance of the harmonized PACC that uses IRT scoring relative to the commonly used standardized PACC that uses z scores. We hypothesize that the harmonized latent PACC (lPACC) scores may show differences in lPACC scores across studies, while the standardized PACC (zPACC) by definition will not. For validation, we examined the influence of baseline Aβ status on longitudinal changes in either lPACC or zPACC, with the hypothesis that the lPACC model may show stronger relationships with Aβ status than zPACC. For an additional validation, we examined baseline lPACC and zPACC scores to predict progression to mild cognitive impairment (MCI) or dementia; we hypothesized baseline lPACC scores would explain more variation in the risk of progression compared to the zPACC.

Method

Participants

Three well-characterized observational cohorts of cognitively intact older adults that have previously published a version of the PACC were used for our base model: ADNI ($n = 795$), HABS ($n = 427$), and AIBL ($n = 1{,}490$). We also report demographic characteristics for the baseline prescreening cohort from the A4 study ($n = 4{,}492$); we use this cohort to demonstrate our method for creating a harmonized PACC from the base model. Data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD (current status information at adni-info.org). For AIBL, data were collected by the AIBL study group. AIBL study methodology has been reported previously (Ellis et al., 2009), with most study data publicly available at ida.loni.usc.edu. A4/LEARN study data were accessed via ida.loni.usc.edu. Participants provided written informed consent prior to study procedures (Aisen et al., 2010; Dagley et al., 2017; Ellis et al., 2009; Sperling et al., 2020). Study protocols were approved by each institutional review board (Massachusetts General Brigham and Austin Health).

Inclusion criteria for each study have been published previously (Aisen et al., 2010; Dagley et al., 2017; Ellis et al., 2009; Sperling et al., 2020). Briefly, we included all participants who were determined by neuropsychologists to be cognitively normal (CN) upon enrollment. Of note, AIBL's initial recruitment was enriched for apolipoprotein (*APOE*) ε4 carriers (Ellis et al., 2009). We used the baseline prescreening data set from the A4 clinical trial cohort, in which individuals were recruited based on their increased risk of cognitive decline and subjective memory concerns (Sperling et al., 2014). The prescreening cohort was not yet screened for Aβ abnormality. The specific inclusion criteria for each study's CN sample are as follows: ADNI: Mini-Mental State Examination (MMSE) = 24–30, logical memory delayed recall (LMDR) ≥ 9 for 16+ years of education, ≥5 for 8–15 years of education, ≥3 for 0–7 years of education, clinical dementia rating (CDR) = 0; HABS: MMSE = 25–30, CDR = 0, education-adjusted LMDR cutoffs equivalent to ADNI; AIBL: MMSE = 26–30, CDR = 0, LMDR cutoffs equivalent to ADNI; A4: MMSE = 25–30, CDR = 0, LMDR = 6–18. HABS and ADNI administered neuropsychological visits annually, while in AIBL, visits were at 18-month intervals. The average number of years and follow-up visits for each cohort are as follows: ADNI = 2.9 (*SD* 3.0) years with 3.7 (3.0) time points; HABS = 4.6 (3.1) years with 5.3 (3.0) time points; AIBL = 4.3 (3.1) years with 2.8 (2.0) time points. Only baseline PACC data were available for the A4 study. In ADNI, HABS, and AIBL, there was a subset who had only baseline data available: ADNI = 286; HABS = 82; AIBL = 242. For each study, a subset of the included participants had available *APOE* genotype data or an Aβ scan (see Table 1). This study was not preregistered and did not have a preregistered analysis plan.

Preclinical Alzheimer Cognitive Composites

For ADNI, HABS, and AIBL, we used the PACC-5 composite, which includes a semantic processing test (Papp et al., 2017). A4, however, does not include this component, and so only the standard version of the PACC (Donohue et al., 2014) was examined. All neuropsychological tests in the PACC for each cohort are detailed in Table 2.

Table 1
Cohort Demographic Characteristics

Cohort	Dataset	*N*	Age (*SD*)	Sex (%*F*)	Aβ status (%+)	APOEε4 (%)	Education	Race/ethnicity (%W/%NH)		MMSE median [range]	LMDR median [range]	zPACC median [range]
ADNI	BL only	277	71.1 (6.4)	63	33	32	16.9 (2.3)	91/95		29 [25–30]	13 [3–21]	0.25 [−1.91–1.41]
	Longitudinal	509	74.2 (5.8)	52	35	30	16.4 (2.6)	91/97	BL	29 [24–30]	13 [5–23]	−0.08 [−2.01–1.59]
									LV	29 [7–30]	13 [0–24]	−0.09 [−10.69–1.76]
HABS	BL only	82	70.8 (8.5)	60	45	23	15.0 (2.9)	79/100		29 [21–30]	13 [0–21]	0.001 [−1.35–1.91]
	Longitudinal	345	71.5 (7.9)	60	28	28	15.9 (2.9)		BL	29 [25–30]	14 [6–25]	0.11 [−2.39–2.06]
								78/98	LV	29 [12–30]	16 [0–24]	−0.02 [−7.92–2.36]
AIBL	BL only	303	72.3 (6.5)	54	35	29	14.7 (2.9)	N/A		28 [24–30]	11 [0–20]	−0.10 [−1.81–1.64]
	Longitudinal	1,176	70.7 (6.7)	58	24	29	14.6 (3.0)		BL	29 [20–30]	11 [0–23]	0.07 [−2.60–1.99]
								N/A	LV	29 [7–30]	12 [0–24]	0.002 [−7.72–1.95]
A4	BL only	4,492	71.3 (4.7)	59	30	33	16.6 (2.8)	91/96		29 [25–30]	12 [4–14]	0.04 [−3.02–1.79]

Note. Aβ = Amyloid status; APOE = apolipoprotein; W = White; NH = Not Hispanic; BL = Baseline; LV = Last visit; MMSE = Mini-Mental State Examination; LMDR = logical memory delayed recall; zPACC = standardized PACC; ADNI = Alzheimer's Disease Neuroimaging Initiative; HABS = Harvard Aging Brain Study; AIBL = Australian Imaging, Biomarker and Lifestyle Study of Ageing; A4 = Anti-Amyloid Treatment of Asymptomatic Alzheimer's Disease; PACC = preclinical Alzheimer cognitive composite.

Table 2
PACC Neuropsychological Test Components Per Cohort

Domain	ADNI	HABS	AIBL	A4
Global	**Total MMSE score**	**Total MMSE score**	**Total MMSE score**	**Total MMSE score**
Story recall memory	**Logical memory delayed recall (Anna Thompson story)**	**Logical memory delayed recall (Anna Thompson story)**	**Logical memory delayed recall (Anna Thompson story)**	Logical memory delayed recall (Robert Miller story)
Executive function	Trails B time	**Digit Symbol Substitution Test (WAIS-R; 90 s)**	Digit Symbol Substitution Test (WAIS-III; 120 s)	**Digit Symbol Substitution Test (WAIS-R; 90 s)**
Verbal fluency	Category fluency—**animals**	Category fluency—**animals** + vegetables + fruits	Category fluency—**animals** + furniture + names	
List learning Memory	ADAS-cog del word recall	**Free and Cued Selective Reminding Test (FCSRT)**	California Verbal Learning Test—2nd Ed. (CVLT-II, long delay)	**Free and Cued Selective Reminding Test (FCSRT)**

Note. Bolded tests reflect cohort shared tests between studies that share recoding distribution and confirmatory factor analysis (CFA) loadings and thresholds (MMSE in all studies; logical memory delayed recall and the animals category in ADNI, HABS, and AIBL; WAIS-R and FCSRT in HABS and A4). Means and standard deviations (*SD*) are provided under the MMSE and logical memory scores for ADNI, AIBL, and HABS to demonstrate differences across the cohorts. PACC = preclinical Alzheimer cognitive composite; ADNI = Alzheimer's Disease Neuroimaging Initiative; HABS = Harvard Aging Brain Study; AIBL = Australian Imaging, Biomarker and Lifestyle Study of Ageing; A4 = Anti-Amyloid Treatment of Asymptomatic Alzheimer's Disease; MMSE = Mini-Mental State Examination; WAIS-R = Wechsler Adult Intelligence Scale-Revised; WAIS-III = Wechsler Adult Intelligence Scale-III; ADAS = Alzheimer's disease Assessment Scale; CVLT = California Verbal Learning Test.

For all cohorts, the standardized method (zPACC) was used as a comparison to the harmonized PACC (lPACC). The zPACC was calculated by standardizing each total test score to its baseline mean and standard deviation and then averaging across these tests at each visit within each cohort. The methods used here for computing the zPACC are aimed to replicate the standardized PACC measure used in previously published analyses (Lim et al., 2016; Mormino et al., 2017; Papp et al., 2017). The tests included in each cohort's zPACC are shown in Table 2.

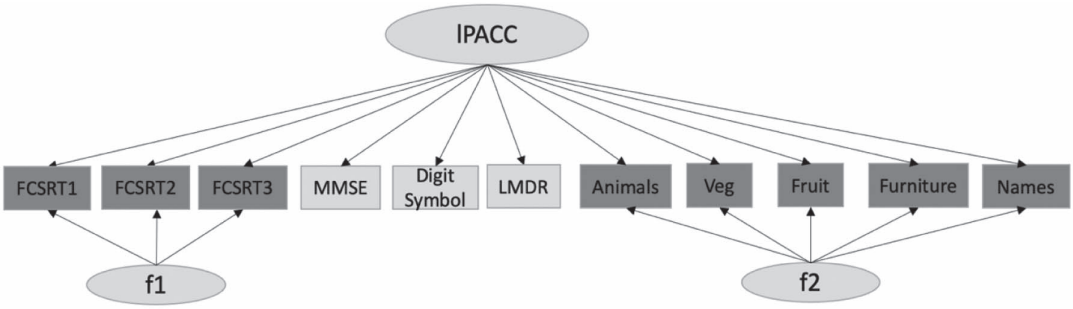
Method for PACC Harmonization

The harmonization approach is a multistep process, which has been detailed in previous publications (Dumitrescu et al., 2020) and summarized here. Using confirmatory factor analysis (CFA), all neuropsychological tests are loaded onto a primary latent factor. Item parameters from all neuropsychological test items are estimated within the base model which includes last visit (LV) data from ADNI, HABS, and AIBL. Cohort-shared tests and items (i.e., MMSE, LMDR, and category fluency—animals) are anchored across studies, which means their item parameters are forced to be the same across studies.

Granular data were available for three trials of the Free and Cued Selective Reminding Test (FCSRT) in HABS and A4 and sets of different category fluency items were available for HABS and AIBL. For HABS and A4 studies, individual trial (granular) data for FCSRT were available, and for AIBL and HABS, granular data were available for multiple category fluency semantics (i.e., raw scores for each of animals, furnitures, and fruits in HABS instead of a total score). Typically, the FCSRT free recall total score is doubled and added to the cued score for the zPACC (Donohue et al., 2014). For the creation of the harmonized PACC, the three trials remained as separate components. The free recall scores were still double weighted within each trial in order to reflect FCSRT literature (Donohue et al., 2014).

We included secondary structures in the base model for those to account for methods effect (FCSRT) and theoretical dependency (category fluency items) to check whether bifactor CFA models provided better fit statistics and stabilized high standardized loadings on some items (see diagrammatic representation in Figure 1).

Figure 1
Bifactor CFA Model Structure at Last Visit

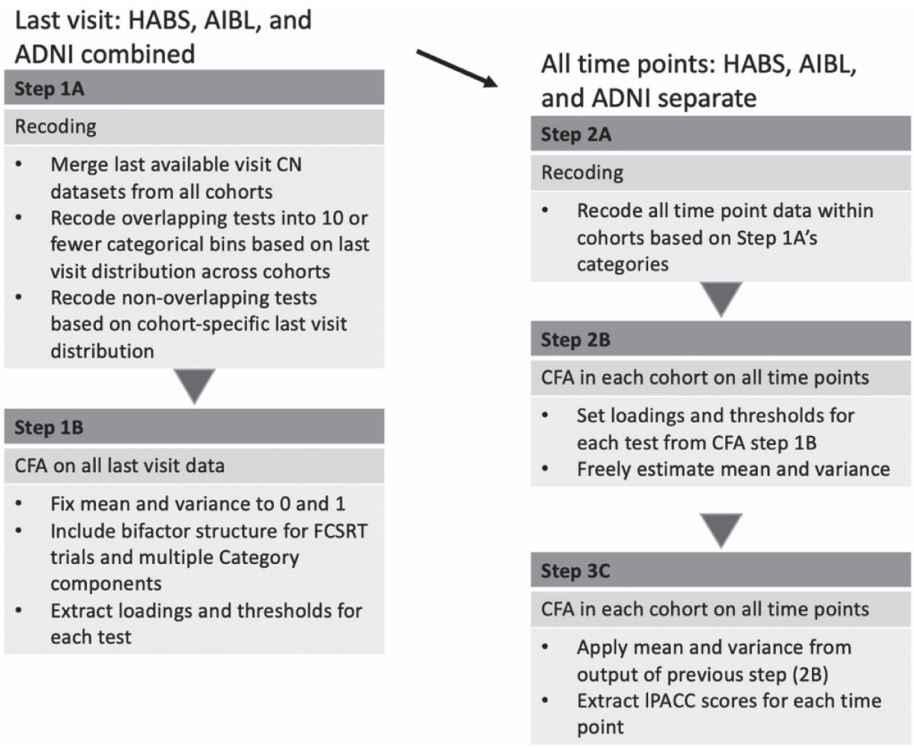


Note. CFA = confirmatory factor analysis; FCSRT = Free and Cued Selective Recall Test trials 1–3; MMSE = Mini-Mental State Examination; LMDR = logical memory delayed recall; lPACC = latent PACC; f1/f2 = secondary latent structures for FCSRT and categories (granular data); each arrow indicates loading of each neuropsychological test onto the latent factors f1, f2, and lPACC.

We selected last visit data from current data pull for our base model to capture greater range in test/item scores. Our full code for the harmonization approach is provided here: https://github.com/rfbuckley/pacc-harmonization. An outline is below and depicted in Figure 2:

1. Recoding raw scores
    a. Raw continuous neuropsychological tests that have scores with a range greater than 10 are recoded into a maximum of 10 categorical bins based on their distribution at the last visit so that they can be used as categorical items in Mplus (Muthén & Muthén, 2017). These bins are coded the same across all tests/items that are common across the cohorts. Tests that are unique to each cohort are recoded based on last visit distribution within that cohort (Figure 2 Steps 1A/2A). Although continuous variables can be analyzed in CFA, the use of categorical scores provides factor scores with less bias (Gross, Jones, et al., 2014; Rhemtulla et al., 2012; Proust-Lima et al., 2007) as there is no need to assume a linear relationship between standard scores and the underlying global cognition factor. To avoid the issue of sparseness affecting model estimates, each category of a given item was required to contain at least five observations.
    b. We chose the final visit to derive the loadings and thresholds for the items rather than the baseline visit as there is a greater range in test/item scores as clinically normal individuals progress through a study. This also helps us calibrate the lower thresholds of items so that we do not need to extrapolate out of range at later visits or as individuals progress to MCI or dementia. The main goal for this step is to ensure a maintenance of the tail end distribution and account for nonlinear relationships between the individual tests and the underlying primary factor (Proust-Lima et al., 2007).
2. Three-step CFA on the combined model
    a. Using the CFA model on a data set with all cohorts' (ADNI, HABS, AIBL) final time points; all neuropsychological test scores were loaded on the latent factor using robust maximum-likelihood (MLR) estimation, with the variance on the general factor set to 1, to derive loadings and thresholds of each test/item (Figure 2 Step 1B).
        i. Additionally, for tests with granular data (FCSRT and category fluency items) the variance of each secondary domain was set to 1 (see Figure 1).

Figure 2
Harmonization Workflow Diagram



Note. CFA = confirmatory factor analysis; ADNI = Alzheimer's Disease Neuroimaging Initiative; HABS = Harvard Aging Brain Study; AIBL = Australian Imaging, Biomarker and Lifestyle Study of Ageing; FCSRT = Free and Cued Selective Recall Test; lPACC = latent PACC; CN = cognitively normal.

b. This model was used to estimate item parameters for each item; these parameters were then used to obtain scores for each time point for each participant. Loadings and thresholds from the calibration model were applied to each cohort's longitudinal data set to obtain scores. The mean and variance of the latent factor were freely estimated (Figure 2 Step 2B).

c. Using the prior model's raw item loadings and thresholds, and the latent factor's mean and variance, all parameters were set for the final CFA for each cohort. This final step is recommended to produce more robust estimates. Scores from the latent factor (lPACC) were then subsequently extracted (Figure 2 Step 2C).

Aβ Positron Emission Tomography

HABS and AIBL acquire ^{11}C-Pittsburgh Compound-B (PiB), positron emission tomography (PET) data, while ADNI uses the ^{18}F-AV45 (Florbetapir) tracer. ADNI and AIBL acquired Aβ-PET data 50–70 min postinjection, whereas HABS PiB-PET data were acquired 40–60 min after injection. Each study's processing pipeline has been published previously (Dagley et al., 2017; Landau et al., 2012; Rowe et al., 2010). Briefly, all PET data underwent reconstruction and attenuation correction and normalized to an Montreal Neurological Institute (MNI) template using SPM12. ADNI used whole cerebellum as a reference region, while AIBL and HABS used cerebellar gray matter as a reference. ADNI and AIBL reported standardized uptake value ratios (SUVRs), while HABS uses a distribution value ratio (DVR) of the frontal, lateral, and retrosplenial regions (FLR). Published cohort-derived cutoffs for high Aβ burden are as follows: HABS: >1.185 DVR (Buckley et al., 2018); AIBL: >1.40 SUVR (Rowe et al., 2010); ADNI: >1.11 SUVR (Landau et al., 2012).

Magnetic Resonance Imaging: Adjusted Hippocampal Volume

All studies acquired 3T Magnetization Prepared Rapid Gradient Echo Imaging (MPRAGE) structural magnetic resonance imaging (MRI). Details regarding MRI data acquisition have been previously outlined (Aisen et al., 2010; Dagley et al., 2017; Ellis et al., 2009). For all cohorts, adjusted hippocampal volume (HV) was calculated using the following algorithm (Mormino et al., 2014):

$$\beta = \text{Beta coefficient bihemispheric HV regressed onto intracranial volume}$$
$$\text{Adjusted HV} = \text{Bihemispheric HV} - (\beta \times (\text{ICV} - \text{sample mean ICV})).$$

Statistical Analyses

A series of CFAs were performed using the Mplus software (Version 8, Muthén & Muthén, Los Angeles CA). Figure 1 presents a graphical depiction of the approach. The choice of running a bifactor versus single-factor CFA was made by directly comparing the bifactor and single models in just the HABS cohort and using a weighted least squares with mean and variance adjusted (WLSMV) estimator. The rationale for the WLSMV estimator is that fit indices are possible to compare between the models. In the full model that involved all cohorts, however, all CFAs were run using a MLR estimator, which does not generate model fit statistics. MLR estimator was used due to the way we defined our base model, since this estimator can account for missingness from neuropsychological tests that are available in one cohort, but not another. Standardized item parameter estimates from all neuropsychological tests/items from the base CFA model, means and variances of primary factor, and standard errors of measurement (SEM) along the estimated latent trait (lPACC) scores are shown (Figure 3G).

We used *R* Version 3.6.3 for all comparisons between the lPACC and the zPACC scores. Correlations between the lPACC and zPACC scores were calculated within each cohort at baseline. We then analyzed demographic and HV relationships with both baseline lPACC and zPACC within each cohort. To describe similarities or differences in how the lPACC and zPACC change over time, we extracted best linear unbiased prediction (BLUP) slopes using linear mixed-effect models including random effects of intercept and time as an estimate of rates of change in the lPACC and zPACC. As the relationship between baseline Aβ status and PACC decline has been extensively reported (Burnham et al., 2016; Buckley et al., 2018; Donohue, Sperling, et al., 2017; Lim et al., 2016; Mormino et al., 2017; Papp et al., 2017, 2020), we performed longitudinal analyses with all cohorts in one model using a linear mixed model to examine the influence of baseline Aβ status on either the lPACC or zPACC over time. Models were adjusted for age, sex, education, and study over time with random intercept and slope in the model to replicate previous publications' analyses. Lastly, we performed a Cox proportional hazards model to compare the lPACC and zPACC in their association with rates of progression to MCI or dementia while covarying for baseline age, sex, education, *APOE* genotype, Aβ status, and cohort. For all analyses, we compared effect sizes to ascertain differences between the lPACC and zPACC.

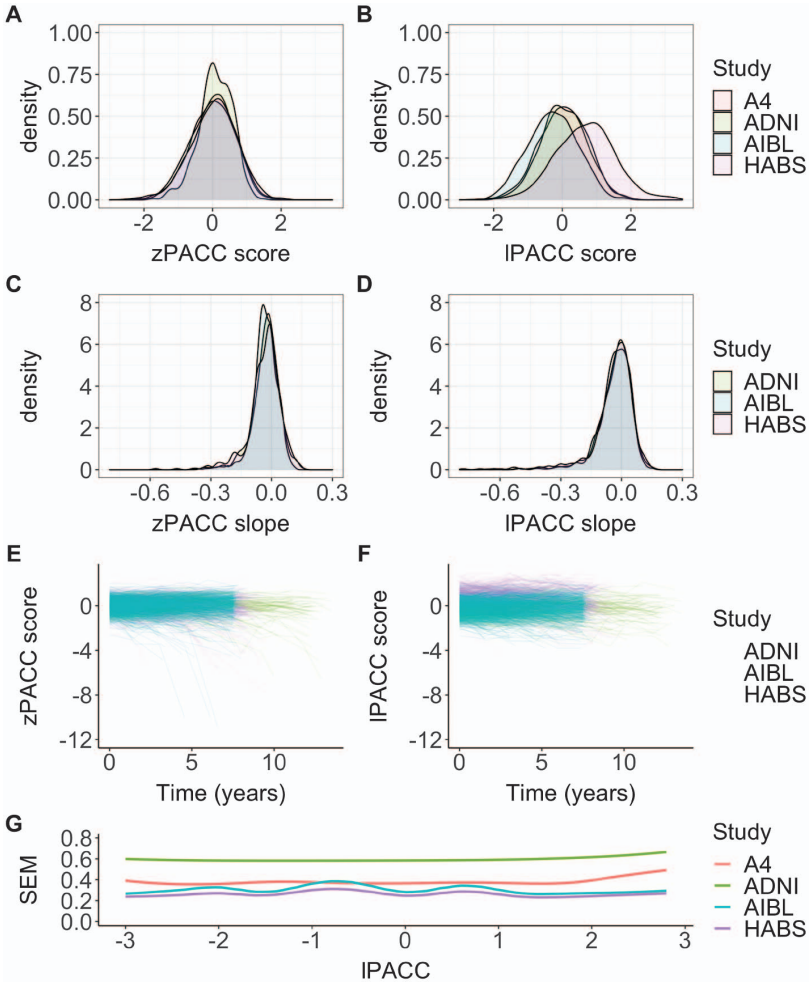
Application of Harmonized Loadings to A4 Cohort

To demonstrate the ability of our method to extract harmonized PACC scores from new data sets, we applied the loadings from our main model to baseline data from the A4 cohort. From this, it can be demonstrated that (a) modern psychometric methods can be used to extend to a new study and (b) the mean and standard deviation in a new sample can be evaluated on the same metric to determine whether the *z*-score assumption of equal means across all cohorts is appropriate for the A4 study.

We performed a similar stepwise CFA on the A4 data set. First, we leveraged neuropsychological tests that overlapped with our base model (MMSE, Wechsler Adult Intelligence Scale-Revised [WAIS-R], FCSRT; Table 2). With these overlapping tests, we set the loadings for all latent structures from the loadings from the base model. We also recoded the overlapping tests based on the base model's categorical bins. An important detail is regarding the recording to categorical bins; if some category bins were missing for overlapping/anchor items in A4 study, the threshold parameters were adjusted accordingly. We then repeated the same CFA methods (Three-step CFA on the combined model section). To determine thresholds and loadings of the unique item(s) in A4, we allowed the unique tests to load onto the latent factor with the mean and variance

Figure 3

zPACC and LPACC Score Distributions Across Cohorts



Note. The left column displays zPACC scores, and the right column displays lPACC scores. (A, B) The baseline distribution of the scores by cohort, (C, D) longitudinal slopes (extracted from linear mixed-effects model) by cohort, and (E, F) a spaghetti plot of PACC performance per participant by cohort. The lPACC scores are shifted from zero at the baseline and show a constrained variance over time compared to the zPACC scores, and (G) smoothed estimates of standard errors of measurement (inverse square root of the total information across test items in each cohort) as a function of the estimated latent trait (from −3 to 3). PACC = preclinical Alzheimer cognitive composite; lPACC = latent PACC; zPACC = standardized PACC; SEM = standard error of measurement; ADNI = Alzheimer's Disease Neuroimaging Initiative; HABS = Harvard Aging Brain Study; AIBL = Australian Imaging, Biomarker and Lifestyle Study of Ageing; A4 = Anti-Amyloid Treatment of Asymptomatic Alzheimer's Disease.

allowed freely estimated (only a single loading and threshold need to be specified for identification). Once all loadings and item thresholds were determined, we ran two CFA models with those set parameters. In the first CFA, the factor's mean and variance were freely estimated, and the final CFA, we set the resulting factor's mean and variance and extracted the lPACC scores for the A4 cohort. We report the model loadings, mean, variance, and SEM. Code and other materials for conducting PACC harmonization can be found here: https://github.com/rfbuckley/pacc-harmonization. Processed lPACC data are available upon request and approval by each respective cohort.

Results

Single-Factor Versus Bifactor Comparison

We ran initial models only in the HABS cohort to examine the fit of the single factor versus a bifactor approach. Using fit statistics of the comparative fit index; CFI (which ranges 0–1, with higher values indicating better fit; values >0.95 are consistent with good fit), Tucker–Lewis index; TLI (which ranges 0–1, with higher values indicating better fit; values >0.95 are consistent with good fit), and root mean square error of approximation; RMSEA (which has a

lower bound of zero and lower values indicate better fit; values <0.08 are adequate fit and <0.06 is excellent fit; Hu & Bentler, 1999), we found that the bifactor model fit better than the single factor ($CFI_{single} = 0.92/CFI_{bifactor} = 0.98$, $TFI_{single} = 0.89/TFI_{bifactor} = 0.97$, $RMSEA_{single} = 0.18/RMSEA_{bifactor} = 0.09$). We also found that 142 last visit data points (33%) had an absolute difference between the models of >0.3 in their lPACC scores using single versus bifactor modeling. Briefly, if scores under the two methods are trivially different from each other, there is no need for the increased computational complexity of the bifactor model. We operationalize "trivially different" as <5% of scores being different by at least 0.3 units. The 0.3 units is the default stopping rule for computerized adaptive tests. In this instance, the choice of a bifactor model was clear on multiple bases—all three fit statistics for the single-factor model were better for the bifactor model, and well over 5% of people had score differences of greater than 0.3. Based on these findings, we opted for bifactor scoring for the lPACC.

Outputs of CFA Statistics in Each Cohort Based on Last Visit Loadings

Full demographic information for each cohort sample is shown in Table 1. The component loadings from the CFA on the last visit data from all cohorts are displayed in the "last visit" section of Table 3. Notably, different indicators in the same category had different loadings (i.e., categories animals, vegetables, and fruits). The longitudinal lPACC score means, variances, and average SEMs are found in Table 3. It is important to note that the lPACC means and variances were very different across the cohorts. This supports the notion that each tests' loadings onto the PACC within each cohort reflects very different latent structures. The SEM is a continuous function across the range of cognitive ability and is proportional to the inverse square root of the total information function of the combined tests within each cohort. A higher SEM at certain scores indicates poorer measurement in that range. Figure 3G shows the smooth estimates of SEM across the estimated latent trait, lPACC. SEMs were higher for ADNI, indicating poorer measurement precision relative to the other cohorts, while A4 showed slightly poorer measurement precision at higher lPACC scores.

lPACC and zPACC Correlate Similarly With Demographics at Baseline

Across all studies, the baseline distributions of both the lPACC and zPACC were normal and displayed similar score variance around the mean. However, the baseline distributions of the lPACC for HABS and AIBL were shifted slightly above and below 0, respectively (Figure 3B). The intercept lPACC and zPACC median and range was 0.113 [−2.33, 3.31] and 0.028 [−2.60, 2.06], respectively. Importantly, while the zPACC, by definition, had mean and standard deviations of approximately 0 and 1, respectively, the mean and standard deviations for the lPACC were vastly different by cohort, $ADNI_{mean(SD)} = 0.23(0.6)$, $AIBL_{mean(SD)} = -0.21(0.7)$, $HABS_{mean(SD)} = 1.05(0.8)$.

We then examined demographic and HV correlations at baseline using lPACC and zPACC within each cohort as a metric of external validity (Table 4). The lPACC and zPACC were associated with age and adjusted HV similarly. The lPACC associations with education were generally weaker than for zPACC, and lPACC scores were

Table 3

Standardized Component Loadings, Unstandardized Mean/Variance, and Standard Errors of Measurement From All Cohorts Last Visits and Cohort-Specific Longitudinal CFAs

Model	Neuropsychological test/item	ADNI	HABS	AIBL	A4
	N	795	427	1,176	4,492
	Std loading				
Last visit model	MMSE	0.559[a]	0.559[a]	0.559[a]	0.559[a]
	LMDR	0.633[a]	0.633[a]	0.633[a]	—
	LMDR_A4	—	—	—	0.281
	Digit symbol	—	0.717[b]	—	0.717[b]
	Digit symbol AIBL	—	—	0.659	—
	FCSRT trial 1	—	0.646[b]	—	0.646[b]
	FCSRT trial 2	—	0.683[b]	—	0.683[b]
	FCSRT trial 3	—	0.620[b]	—	0.620[b]
	Trails B	0.608	—	—	—
	ADAS-cog	0.596	—	—	—
	Animals	0.588[+]	0.588[a]	0.588[a]	—
	Vegetables	—	0.706	—	—
	Fruits	—	0.695	0.695	—
	Names	—	—	0.658	—
	Furniture	—	—	0.676	—
	CVLT	—	—	0.719	—
Longitudinal model	Mean/variance	0.241/0.567	0.823/0.951	−0.312/0.618	0.240/0.431
	SEM mean(*SD*)	0.50 (0.04)	0.28 (0.07)	0.32 (0.05)	0.34 (0.02)

Note. Std = standardized; SEM = standard error of measurement; ADNI = Alzheimer's Disease Neuroimaging Initiative; HABS = Harvard Aging Brain Study; CFA = confirmatory factor analysis; AIBL = Australian Imaging, Biomarker and Lifestyle Study of Ageing; A4 = Anti-Amyloid Treatment of Asymptomatic Alzheimer's Disease; MMSE = Mini-Mental State Examination; LMDR = logical memory delayed recall; FCSRT = Free and Cued Selective Reminding Test; ADAS = Alzheimer's disease Assessment Scale; CVLT = California Verbal Learning Test.
[a] Tests that overlap between cohorts have their loadings locked in the base model. [b] Shared loadings for the secondary model to A4.

Table 4

Correlations of zPACC and lPACC at the Baseline Within Each Cohort Reported Using Pearson's Correlation Coefficients and t-Test Effect Size Statistics

		ADNI		HABS		AIBL		A4	
Variable	Estimate	zPACC	lPACC	zPACC	lPACC	zPACC	lPACC	zPACC	lPACC
	r	0.96*		0.83*		0.77*		0.81*	
Education	*r*	0.27*	0.27*	0.18**	0.15***	0.29*	0.18*	0.15*	0.06*
Age	*r*	−0.37*	−0.38*	−0.36*	−0.31*	−0.40*	−0.35*	−0.31*	−0.30*
Sex	*d*	0.18**	0.19**	0.40*	0.57*	0.42*	0.51*	0.53*	0.58*
HVadj	*r*	0.26*	0.25*	0.24*	0.21*	0.26*	0.19*	0.31*	0.28*

Note. HVadj = hippocampal volume adjusted for intracranial volume; ADNI = Alzheimer's Disease Neuroimaging Initiative; HABS = Harvard Aging Brain Study; AIBL = Australian Imaging, Biomarker and Lifestyle Study of Ageing; A4 = Anti-Amyloid Treatment of Asymptomatic Alzheimer's Disease; PACC = preclinical Alzheimer cognitive composite; lPACC = latent PACC; zPACC = standardized PACC.
* $p < .0001$. ** $p < .001$. *** $p < .01$.

much more pronounced between the sexes in HABS and AIBL relative to the zPACC (Table 4).

Longitudinal lPACC Performs Similarly in a Linear Mixed Model With Aβ Status to zPACC

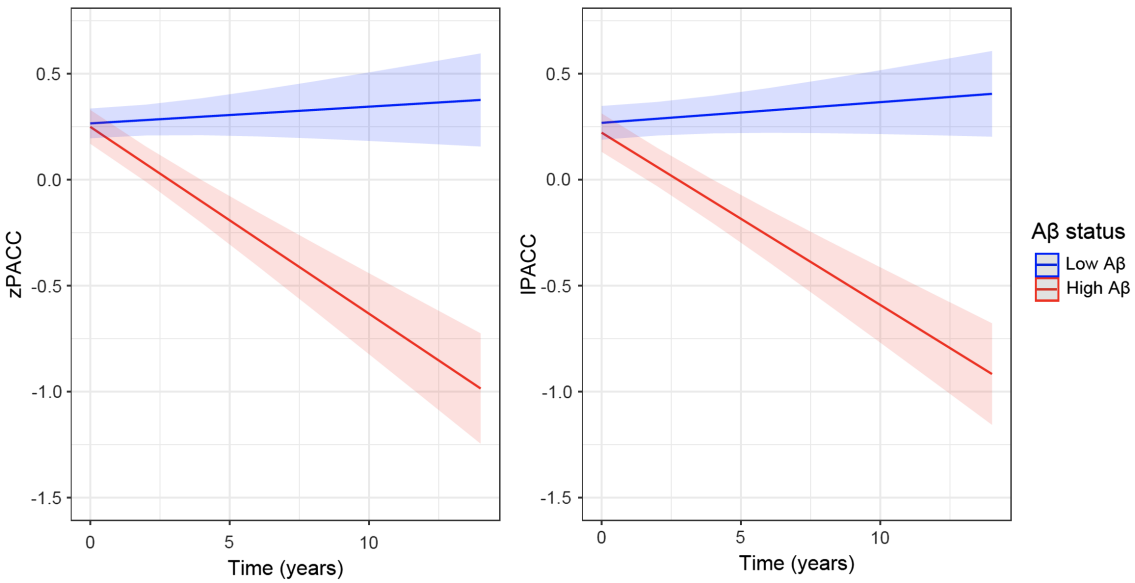
Extracted lPACC slopes were normally distributed across all the cohorts (see Figure 3C), with slightly more constricted variance around the mean compared to the zPACC score. The zPACC slopes showed outliers in the negative tail (Figure 3C), as can be demonstrated in the median and range of the distributions: lPACC slope median and range: −0.036 [−0.519, 0.124]; zPACC slope median and range: −0.021 [−0.876, 0.154]. We merged ADNI, HABS, and AIBL longitudinal data to observe how each version of the PACC performed over time as influenced by baseline Aβ status using linear mixed models. The *t* value for Aβ status on lPACC change was larger than that of the longitudinal zPACC: lPACC: $t(6,978) = -10.43$, $SE = 0.006$, $p < .001$; zPACC: $t(6,978) = -9.89$, $SE = 0.006$, $p < .001$ (Figure 4).

Cox Proportional Hazards Model

We performed a survival analysis with all cohorts using baseline amyloid status, sex, age, years of education, *APOE* genotype, and either lPACC or zPACC to predict progression to MCI or dementia. Each model was fitting to 181 events of progression to MCI or dementia. Both the zPACC and lPACC performed similarly; however, the lPACC slightly outperformed the zPACC. The hazard ratios (HR) for the lPACC were HR(95% confidence interval [CI]) = 0.491(0.386–0.626), $p < .0001$, and for zPACC were, HR(95% CI) = 0.424 (0330–0.546), $p < .0001$. The pseudo-R^2 for the lPACC model was higher at 58% variance explained in the model, relative to 56% for the zPACC model.

Figure 4

Baseline Aβ Status on Longitudinal PACC Change With All Cohorts



Note. High Aβ is in red, while low Aβ is in blue. The lPACC (right panel) compared with zPACC (left panel). PACC = preclinical Alzheimer cognitive composite; lPACC = latent PACC; zPACC = standardized PACC.

Harmonizing Base Model to A4 Cohort

The purpose of this section is to demonstrate how other cohorts can be harmonized to the base model. We used the overlapping MMSE scores with the other three cohorts, and the overlapping FCSRT and digit symbol scores with HABS. The loading for the unique test (LMDR Robert Miller Story), the lPACC mean and variance, and the standard error of means for the A4 CFAs are reported in Table 3. We found baseline A4 lPACC scores were distributed similarly to ADNI lPACC scores (Table 3). Demographic and adjusted HV correlations between lPACC and zPACC within A4 were consistent with patterns seen in HABS and AIBL (Table 4).

Discussion

In this work we presented a method to harmonize the PACC across three well-characterized cohorts of clinically unimpaired older individuals, followed by demonstrating an extension of our approach to a fourth similar cohort. This is a more sophisticated approach to cognitive data harmonization than standardization that captures the essence of the contribution of specific cognitive domains through IRT. The initial intention of the PACC was to combine multiple tests that changed significantly in clinically normal older adults with elevated amyloid. The IRT approach used in deriving the lPACC allows for a more direct quantification of this global composite. Overall, the procedure performed as expected compared to previous applications of this method with other composites (Dumitrescu et al., 2020). Examining estimated SEMs, we found measurement precision to be poorer for ADNI relative to other cohorts, and HABS and AIBL to be the highest. The latter had the most test items available for the bifactor model. We were able to recapitulate cross-sectional findings from the zPACC while revealing performance differences between cohorts at the intercept. By definition, the zPACC will look similar across all the cohorts at the intercept because they are forced to a normal (0, 1) distribution using the standardization approach. Notably, however, the means and standard deviations of the lPACC scores at the intercept were markedly different, highlighting the inherent differences between the cohorts in these composites potentially arising from recruitment priorities idiosyncratic to each cohort. As such, the assumption of the standardization approach, that all composites are equally weighted and abide by the same distributional properties, is clearly and impressively violated in this case. This was also demonstrated by the different means and standard deviations for the overlapping tests (MMSE and LMDR) presented in Table 2, suggesting that an individuals score on the zPACC is dependent on the sample in which the individual comes from, even if the underlying scores are exactly the same. By contrast, using IRT approaches, such as the one we present, will produce a factor score that is considerably less dependent on the cohort than the standardization approach. That is, item responses will be scored similarly across the study/cohort/country the individual is tested in. Indeed, it seems clear that the global mean score of the zPACC is obscuring critical cross-sectional study differences that should be acknowledged. This further supports our argument that modern psychometric approaches to harmonizing cognitive data provide more flexible parameterization that does not require the vast assumptions of the *z*-score approach.

There are some important decisions made in the CFA approach that requires highlighting. The last visit was chosen to form the basis of this harmonization method. The rationale was to capture a greater between-subject score variability in the extreme ends of the distribution. A specific advantage to using the last visit distribution is to minimize overinflated estimates of decline over time based on the constrained variance available at the baseline, particularly regarding the MMSE in a CN sample. As shown in Figure 3D and E, the lPACC score range was constrained and resulted in less curvilinear decline relative to the zPACC. This decision could be applied to the zPACC but has typically not been used in previous studies. Another potential rationale for the constrained variance in the lPACC over time could be due to a loss of information when converting the continuous test scores into categorical in the rescoring step. We chose to categorize the test data to produce less bias in the resulting factor scores (Gross, Sherva, et al., 2014; Proust-Lima et al., 2007; Rhemtulla et al., 2012); however, it is possible that this may impact calibration of data at the tail ends of the distribution of scores.

Another important decision is the selection of a baseline CN cohort for our method as it limits the application of the lPACC beyond CN participants to those with cognitive impairment. The PACC was intended for detection of Aβ-related change in a CN sample (Donohue et al., 2014), and so we aimed to develop a harmonization approach that focused on score distributions within this group. Examinations of the lPACC or zPACC in diagnosed patients (i.e., MCI or dementia) using scores calibrated on clinically normal adults is problematic due to the need to extrapolate beyond the distribution deriving the metric. To avoid the issue of constricted score ranges in CN individuals, we have opted for a large cohort that exhibits a full range of cognitive scores at the last visit to ensure that a floor effect does not occur in the estimated scores. Regardless, it is important to note that our sample has a constrained distribution relative to MCI and dementia, even at the last visit, which will result in thresholds and loadings that would be unrealistic for a clinically impaired sample and may underestimate change in these groups. Future studies requiring harmonized PACC in clinically impaired samples would need to calibrate scores based on data from a sample with greater cognitive range.

Further, the harmonization method applied here used item-level (granular) data, which is markedly different from the standardized approach, which only uses total scores from each test. The rationale behind using granular data is that total summed scores make important assumptions about the weight of each item and what each item contributes to the total score. A second issue is related to curvilinearity: Standard scores make important (and often untested) assumptions that the resulting score has linear scaling properties. This is not always the case, as has been demonstrated with the MMSE (Crane et al., 2008; Lopez et al., 2005; Proust-Lima et al., 2007; Tombaugh & McIntyre, 1992; Wind et al., 1997). Inclusion of as much granular data as available is preferred to calculate a model that captures sources of covariation that we expect are operating in the data such as methods effects from multiple list learning trials. As computing infrastructure and digital data collection become more ubiquitous in the clinic and in research studies, it will become increasingly possible to capture and use granular data, which could also be summed into total scores if assumptions of those scores are met. In many cases, however, those assumptions do not hold, and the granular data permit modeling that better reflects the ability levels of the participant. In cases where only total scores are available, it is still possible to harmonize with the base model, but there will be some drawbacks. For instance, if category fluency granular data is unavailable, the total score can be used but not anchored to the other

cohorts. On a positive note, however, even if only one test is overlapping with the other cohorts, it is possible to harmonize the PACC to the base model.

We observed impressive cohort score differences at the baseline using the lPACC, potentially highlighting differences between the cohorts, like recruitment differences, that a standardized composite such as the zPACC obscures. Additionally, when A4 data were harmonized to the base model, we observed that loading differences were different for logical memory. The A4 study uses the Robert Miller version of logical memory, while the base cohorts use the Anna Thompson version of logical memory. Previous analyses have shown discrepancies in difficulty between story versions based on their sentence length and grammatical complexity (Morris et al., 1997). This further exemplifies how test version and administration variation can differentially influence composite scores and should not be weighted equally across cohorts unless that equal weighting can be confirmed. The modern psychometric approach enables us to formally test whether cognitive testing is equally difficult. And as in the present case when the two stimuli have very different parameters, the flexible IRT approach enables us to treat these as separate items, still enabling resulting scores to be on the same metric while accounting for these differences in item levels. It is also possible, however, that other factors may contribute to such a different loading structure for logical memory in A4: First and foremost, A4 uses a four-test composite to define the PACC, which is unlike the other cohorts that use a five-test composite. It is possible that the A4 lPACC scores are less precise due to fewer components contributing to the overall variance structure. The A4 study is also unique in that only baseline neuropsychological scores were available. The component loadings applied from the other three cohorts were calibrated to the last available visits, and as such, this difference in overall score variance could also contribute to a low factor loading for A4's logical memory test.

Although less "decline" was evident on the lPACC, we found that lPACC slightly outperformed the zPACC in both the linear mixed models (longitudinal PACC) and survival analyses (baseline PACC). It is important to note, however, that the magnitudes of effect were somewhat similar. A potential explanation for this could be that the lPACC compensates for having fewer outliers by exhibiting constrained longitudinal variance relative to the zPACC, resulting in the maintenance of the overall effect size of the association between baseline Aβ status and PACC change.

It is important to recognize the demographic characteristics of the cohorts in this study; participants in these studies represent a highly educated and predominantly white demographic stratification. We will need to account for differential item functioning when we add diverse cohorts to our harmonization pipeline. This limitation is not unique to this approach and also applies to standardization and other composite measures. Still, in future analyses, we aim to include cohorts that are more racially and educationally diverse to better understand the role of race, ethnicity, education, and economic status on the PACC. Further, there is somewhat limited international representation, and so future work will seek to harmonize data sets from cohorts in other countries. Other limitations of our study relate to the availability of granular-level data; the MMSE may have some differences in administration across the cohorts (i.e., the administration or scoring of WORLD backwards or the orientation questions), but due to the lack of item-level data in many of these cohorts, this latent structure could not be included in the bifactor model. Further, it is possible that the administration of different MMSE versions could affect the final scores (i.e., spelling vs. subtraction scores). We need to examine whether modeling granular-level MMSE data might increase the sensitivity of lPACC performance relative to zPACC.

The PACC harmonization technique produces scores on the same metric irrespective of study specific, test item, or test battery idiosyncrasies. These results present an opportunity to conduct statistically powerful and demographically diverse analyses across multiple cohorts using directly comparable scores. A major advantage to this approach is a reduction in noise that the standardized PACC variants introduce when treating them equally across cohorts. Although our methods limit the application of our specific loadings and thresholds to diagnostic groups, our methods are open source which will enable investigators' own harmonization and/or replication of our results. In summary, this harmonization approach to the PACC allows for cocalibrating across cohorts, thus broadening the opportunity for larger sample analyses involving cohorts of preclinical AD individuals. It is important, however, to acknowledge the low-precision measurement of the lPACC in some cohorts, indicating that this composite may not be appropriate to use in situations that require precision measurement at high scores. Harmonization approaches have been published for domain-specific cognitive composites, such as episodic memory, executive function, language, and visuospatial abilities (Crane et al., 2012; Choi et al., 2020), and so it is upon the researcher to ensure the appropriate cognitive composite for their specific research question. As the zPACC is used in AD clinical trials, such as the A4 study, there is a clear need for a harmonized version of this composite to allow for multicohort analyses that benefit from analyzing research questions with these types of data sets.

References

Aisen, P. S., Petersen, R. C., Donohue, M. C., Gamst, A., Raman, R., Thomas, R. G., Walter, S., Trojanowski, J. Q., Shaw, L. M., Beckett, L. A., Jack, C. R., Jr., Jagust, W., Toga, A. W., Saykin, A. J., Morris, J. C., Green, R. C., Weiner, M. W., & the Alzheimer's Disease Neuroimaging Initiative. (2010). Clinical core of the Alzheimer's disease neuroimaging Initiative: Progress and plans. *Alzheimer's and Dementia*, *6*(3), 239–246. https://doi.org/10.1016/j.jalz.2010.03.006

Betthauser, T. J., Koscik, R. L., Jonaitis, E. M., Allison, S. L., Cody, K. A., Erickson, C. M., Rowley, H. A., Stone, C. K., Mueller, K. D., Clark, L. R., Carlsson, C. M., Chin, N. A., Asthana, S., Christian, B. T., & Johnson, S. C. (2020). Amyloid and tau imaging biomarkers explain cognitive decline from late middle-age. *Brain: A Journal of Neurology*, *143*(1), 320–335. https://doi.org/10.1093/brain/awz378

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511490026

Buckley, R. F., Mormino, E. C., Amariglio, R. E., Properzi, M. J., Rabin, J. S., Lim, Y. Y., Papp, K. V., Jacobs, H. I. L., Burnham, S., Hanseeuw, B. J., Doré, V., Dobson, A., Masters, C. L., Waller, M., Rowe, C. C., Maruff, P., Donohue, M. C., Rentz, D. M., Kirn, D., … the Alzheimer's Disease Neuroimaging Initiative, the Australian Imaging, Biomarker and Lifestyle Study of Ageing, the Harvard Aging Brain Study. (2018). Sex, amyloid, and *APOE* ε4 and risk of cognitive decline in preclinical Alzheimer's disease: Findings from three well-characterized cohorts. *Alzheimer's and Dementia*, *14*(9), 1193–1203. https://doi.org/10.1016/j.jalz.2018.04.010

Burnham, S. C., Bourgeat, P., Doré, V., Savage, G., Brown, B., Laws, S., Maruff, P., Salvado, O., Ames, D., Martins, R. N., Masters, C. L., Rowe, C. C., Villemagne, V. L., & the AIBL Research Group. (2016). Clinical

and cognitive trajectories in cognitively healthy elderly individuals with suspected non-Alzheimer's disease pathophysiology (SNAP) or Alzheimer's disease pathology: A longitudinal study. *Lancet Neurology*, *15*(10), 1044–1053. https://doi.org/10.1016/S1474-4422(16)30125-9

Chan, K. S., Gross, A. L., Pezzin, L. E., Brandt, J., & Kasper, J. D. (2015). Harmonizing measures of cognitive performance across international surveys of aging using item response theory. *Journal of Aging and Health*, *27*(8), 1392–1414. https://doi.org/10.1177/0898264315583054

Chapman, L. J., & Chapman, J. P. (1978). The measurement of differential deficit. *Journal of Psychiatric Research*, *14*(1–4), 303–311. https://doi.org/10.1016/0022-3956(78)90034-1

Chhetri, J. K., de Souto Barreto, P., Cantet, C., Cesari, M., Coley, N., Andrieu, S., & Vellas, B. (2017). Trajectory of the MAPT-PACC-preclinical Alzheimer cognitive composite in the placebo group of a randomized control trial: Results from the MAPT study: Lessons for further trials. *The Journal of Prevention of Alzheimer's Disease*, *5*(1), 31–35. https://doi.org/10.14283/jpad.2017.21

Choi, S.-E., Mukherjee, S., Gibbons, L. E., Sanders, R. E., Jones, R. N., Tommet, D., Mez, J., Trittschuh, E. H., Saykin, A., Lamar, M., Rabin, L., Foldi, N. S., Sikkes, S., Jutten, R. J., Grandoit, E., Mac Donald, C., Risacher, S., Groot, C., Ossenkoppele, R., … the Alzheimer's Disease Neuroimaging Initiative. (2020). Development and validation of language and visuospatial composite scores in ADNI. *Alzheimer's and Dementia: Translational Research and Clinical Interventions*, *6*(1), Article e12072. https://doi.org/10.1002/trc2.12072

Crane, P. K., Carle, A., Gibbons, L. E., Insel, P., Mackin, R. S., Gross, A., Jones, R. N., Mukherjee, S., Curtis, S. M., Harvey, D., Weiner, M., Mungas, D., & the Alzheimer's Disease Neuroimaging Initiative. (2012). Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Brain Imaging and Behavior*, *6*(4), 502–516. https://doi.org/10.1007/s11682-012-9186-z

Crane, P. K., Narasimhalu, K., Gibbons, L. E., Mungas, D. M., Haneuse, S., Larson, E. B., Kuller, L., Hall, K., & van Belle, G. (2008). Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. *Journal of Clinical Epidemiology*, *61*(10), 1018–1027.e9. https://doi.org/10.1016/j.jclinepi.2007.11.011

Dagley, A., LaPoint, M., Huijbers, W., Hedden, T., McLaren, D. G., Chatwal, J. P., Papp, K. V., Amariglio, R. E., Blacker, D., Rentz, D. M., Johnson, K. A., Sperling, R. A., & Schultz, A. P. (2017). Harvard aging brain study: Dataset and accessibility. *NeuroImage*, *144*(Pt B), 255–258. https://doi.org/10.1016/j.neuroimage.2015.03.069

Donohue, M. C., Sperling, R. A., Petersen, R., Sun, C.-K., Weiner, M. W., Aisen, P. S., & the Alzheimer's Disease Neuroimaging Initiative. (2017). Association between elevated brain amyloid and subsequent cognitive decline among cognitively normal persons. *JAMA*, *317*(22), 2305–2316. https://doi.org/10.1001/jama.2017.6669

Donohue, M. C., Sperling, R. A., Salmon, D. P., Rentz, D. M., Raman, R., Thomas, R. G., Weiner, M., Aisen, P. S., & the Australian Imaging, Biomarkers, and Lifestyle Flagship Study of Ageing, the Alzheimer's Disease Neuroimaging Initiative, the Alzheimer's Disease Cooperative Study. (2014). The preclinical Alzheimer cognitive composite: Measuring amyloid-related decline. *JAMA Neurology*, *71*(8), 961–970. https://doi.org/10.1001/jamaneurol.2014.803

Donohue, M. C., Sun, C.-K., Raman, R., Insel, P. S., Aisen, P. S., the North American Alzheimer's Disease Neuroimaging Initiative, the Australian Imaging, Biomarker and Lifestyle Study of Ageing, & the Japanese-Alzheimer's Disease Neuroimaging Initiative. (2017). Cross-validation of optimized composites for preclinical Alzheimer's disease. *Alzheimer's and Dementia: Translational Research and Clinical Interventions*, *3*(1), 123–129. https://doi.org/10.1016/j.trci.2016.12.001

Dowling, N. M., Hermann, B., La Rue, A., & Sager, M. A. (2010). Latent structure and factorial invariance of a neuropsychological test battery for the study of preclinical Alzheimer's disease. *Neuropsychology*, *24*(6), 742–756. https://doi.org/10.1037/a0020176

Dumitrescu, L., Mahoney, E. R., Mukherjee, S., Lee, M. L., Bush, W. S., Engelman, C. D., Lu, Q., Fardo, D. W., Trittschuh, E. H., Mez, J., Kaczorowski, C., Hernandez Saucedo, H., Widaman, K. F., Buckley, R., Properzi, M., Mormino, E., Yang, H.-S., Harrison, T., … Hohman, T. J. (2020). Genetic variants and functional pathways associated with resilience to Alzheimer's disease. *Brain: A Journal of Neurology*, *143*(8), 2561–2575. https://doi.org/10.1093/brain/awaa209

Ellis, K. A., Bush, A. I., Darby, D., De Fazio, D., Foster, J., Hudson, P., Lautenschlager, N. T., Lenzo, N., Martins, R. N., Maruff, P., Masters, C., Milner, A., Pike, K., Rowe, C., Savage, G., Szoeke, C., Taddei, K., Villemagne, V., Woodward, D., & the AIBL Research Group. (2009). The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: Methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *International Psychogeriatrics*, *21*(4), 672–687. https://doi.org/10.1017/S1041610209009405

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189–198. https://doi.org/10.1016/0022-3956(75)90026-6

Gibbons, L. E., Feldman, B. J., Crane, H. M., Mugavero, M., Willig, J. H., Patrick, D., Schumacher, J., Saag, M., Kitahata, M. M., & Crane, P. K. (2011). Migrating from a legacy fixed-format measure to CAT administration: Calibrating the PHQ-9 to the PROMIS depression measures. *Quality of Life Research*, *20*(9), 1349–1357. https://doi.org/10.1007/s11136-011-9882-y

Gross, A. L., Jones, R. N., Fong, T. G., Tommet, D., & Inouye, S. K. (2014). Calibration and validation of an innovative approach for estimating general cognitive performance. *Neuroepidemiology*, *42*(3), 144–153. https://doi.org/10.1159/000357647

Gross, A. L., Power, M. C., Albert, M. S., Deal, J. A., Gottesman, R. F., Griswold, M., Wruck, L. M., Mosley, T. H., Jr., Coresh, J., Sharrett, A. R., & Bandeen-Roche, K. (2015). Application of latent variable methods to the study of cognitive decline when tests change over time. *Epidemiology*, *26*(6), 878–887. https://doi.org/10.1097/EDE.0000000000000379

Gross, A. L., Sherva, R., Mukherjee, S., Newhouse, S., Kauwe, J. S. K., Munsie, L. M., Waterston, L. B., Bennett, D. A., Jones, R. N., Green, R. C., Crane, P. K., & the Alzheimer's Disease Neuroimaging Initiative, the GENAROAD Consortium, the AD Genetics Consortium. (2014). Calibrating longitudinal cognition in Alzheimer's disease across diverse test batteries and datasets. *Neuroepidemiology*, *43*(3–4), 194–205. https://doi.org/10.1159/000367970

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Insel, P. S., Weiner, M., Mackin, R. S., Mormino, E., Lim, Y. Y., Stomrud, E., Palmqvist, S., Masters, C. L., Maruff, P. T., Hansson, O., & Mattsson, N. (2019). Determining clinically meaningful decline in preclinical Alzheimer disease. *Neurology*, *93*(4), e322–e333. https://doi.org/10.1212/WNL.0000000000007831

Jessen, F., Spottke, A., Boecker, H., Brosseron, F., Buerger, K., Catak, C., Fliessbach, K., Franke, C., Fuentes, M., Heneka, M. T., Janowitz, D., Kilimann, I., Laske, C., Menne, F., Nestor, P., Peters, O., Priller, J., Pross, V., Ramirez, A., … Düzel, E. (2018). Design and first baseline data of the DZNE multicenter observational study on predementia Alzheimer's disease (DELCODE). *Alzheimer's Research and Therapy*, *10*(1), Article 15. https://doi.org/10.1186/s13195-017-0314-2

Laiacona, M., Barbarotto, R., & Capitani, E. (1998). Semantic category dissociations in naming: Is there a gender effect in Alzheimer's disease? *Neuropsychologia*, *36*(5), 407–419. https://doi.org/10.1016/S0028-3932(97)00125-5

Landau, S. M., Mintun, M. A., Joshi, A. D., Koeppe, R. A., Petersen, R. C., Aisen, P. S., Weiner, M. W., Jagust, W. J., & the Alzheimer's Disease

Neuroimaging Initiative. (2012). Amyloid deposition, hypometabolism, and longitudinal cognitive decline. *Annals of Neurology*, *72*(4), 578–586. https://doi.org/10.1002/ana.23650

Langa, K. M., Ryan, L. H., McCammon, R. J., Jones, R. N., Manly, J. J., Levine, D. A., Sonnega, A., Farron, M., & Weir, D. R. (2020). The health and retirement study harmonized cognitive assessment protocol (HCAP) project: Study design and methods. *Neuroepidemiology*, *54*(1), 64–74. https://doi.org/10.1159/000503004

Lim, Y. Y., Snyder, P. J., Pietrzak, R. H., Ukiqi, A., Villemagne, V. L., Ames, D., Salvado, O., Bourgeat, P., Martins, R. N., Masters, C. L., Rowe, C. C., & Maruff, P. (2016). 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. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, *2*(1), 19–26. https://doi.org/10.1016/j.dadm.2015.11.003

Lopez, M. N., Charter, R. A., Mostafavi, B., Nibut, L. P., & Smith, W. E. (2005). Psychometric properties of the folstein mini-mental state examination. *Assessment*, *12*(2), 137–144. https://doi.org/10.1177/1073191105275412

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*(6), 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Mormino, E. C., Betensky, R. A., Hedden, T., Schultz, A. P., Amariglio, R. E., Rentz, D. M., Johnson, K. A., & Sperling, R. A. (2014). Synergistic effect of β-amyloid and neurodegeneration on cognitive decline in clinically normal individuals. *JAMA Neurology*, *71*(11), 1379–1385. https://doi.org/10.1001/jamaneurol.2014.2031

Mormino, E. C., Papp, K. V., Rentz, D. M., Donohue, M. C., Amariglio, R., Quiroz, Y. T., Chhatwal, J., Marshall, G. A., Donovan, N., Jackson, J., Gatchel, J. R., Hanseeuw, B. J., Schultz, A. P., Aisen, P. S., Johnson, K. A., & Sperling, R. A. (2017). Early and late change on the preclinical Alzheimer's cognitive composite in clinically normal older individuals with elevated amyloid β. *Alzheimer's and Dementia*, *13*(9), 1004–1012. https://doi.org/10.1016/j.jalz.2017.01.018

Morris, J., Kunka, J. M., & Rossini, E. D. (1997). Development of alternate paragraphs for the logical memory subtest of the Wechsler Memory Scale-Revised. *The Clinical Neuropsychologist*, *11*(4), 370–374. https://doi.org/10.1080/13854049708400465

Mukherjee, S., Mez, J., Trittschuh, E. H., Saykin, A. J., Gibbons, L. E., Fardo, D. W., Wessels, M., Bauman, J., Moore, M., Choi, S.-E., Gross, A. L., Rich, J., Louden, D. K. N., Sanders, R. E., Grabowski, T. J., Bird, T. D., McCurry, S. M., Snitz, B. E., … Crane, P. K. (2020). Genetic data and cognitively defined late-onset Alzheimer's disease subgroups. *Molecular Psychiatry*, *25*(11), 2942–2951. https://doi.org/10.1038/s41380-018-0298-8

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th Ed.) [Computer software]. https://github.com/rfbuckley/pacc-harmonization

Papp, K. V., Buckley, R., Mormino, E., Maruff, P., Villemagne, V. L., Masters, C. L., Johnson, K. A., Rentz, D. M., Sperling, R. A., Amariglio, R. E., & the Collaborators from the Harvard Aging Brain Study, the Alzheimer's Disease Neuroimaging Initiative and the Australian Imaging, Biomarker and Lifestyle Study of Aging. (2020). Clinical meaningfulness of subtle cognitive decline on longitudinal testing in preclinical AD. *Alzheimer's and Dementia*, *16*(3), 552–560. https://doi.org/10.1016/j.jalz.2019.09.074

Papp, K. V., Rentz, D. M., Orlovsky, I., Sperling, R. A., & Mormino, E. C. (2017). Optimizing the preclinical Alzheimer's cognitive composite with semantic processing: The PACC5. *Alzheimer's and Dementia: Translational Research and Clinical Interventions*, *3*(4), 668–677. https://doi.org/10.1016/j.trci.2017.10.004

Park, L. Q., Gross, A. L., McLaren, D. G., Pa, J., Johnson, J. K., Mitchell, M., Manly, J. J., & the Alzheimer's Disease Neuroimaging Initiative. (2012). Confirmatory factor analysis of the ADNI Neuropsychological Battery. *Brain Imaging and Behavior*, *6*(4), 528–539. https://doi.org/10.1007/s11682-012-9190-3

Proust-Lima, C., Amieva, H., Dartigues, J.-F., & Jacqmin-Gadda, H. (2007). Sensitivity of four psychometric tests to measure cognitive changes in brain aging-population-based studies. *American Journal of Epidemiology*, *165*(3), 344–350. https://doi.org/10.1093/aje/kwk017

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. https://doi.org/10.1037/a0029315

Rowe, C. C., Ellis, K. A., Rimajova, M., Bourgeat, P., Pike, K. E., Jones, G., Fripp, J., Tochon-Danguy, H., Morandeau, L., O'Keefe, G., Price, R., Raniga, P., Robins, P., Acosta, O., Lenzo, N., Szoeke, C., Salvado, O., Head, R., Martins, R., … Villemagne, V. L. (2010). Amyloid imaging results from the Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging. *Neurobiology of Aging*, *31*(8), 1275–1283. https://doi.org/10.1016/j.neurobiolaging.2010.04.007

Schneider, L. S., & Goldberg, T. E. (2020). Composite cognitive and functional measures for early stage Alzheimer's disease trials. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, *12*(1), Article e12017. https://doi.org/10.1002/dad2.12017

Sperling, R. A., Donohue, M. C., Raman, R., Sun, C.-K., Yaari, R., Holdridge, K., Siemers, E., Johnson, K. A., Aisen, P. S., & the A4 Study Team. (2020). Association of factors with elevated amyloid burden in clinically normal older individuals. *JAMA Neurology*, *77*(6), 735–745. https://doi.org/10.1001/jamaneurol.2020.0387

Sperling, R. A., Rentz, D. M., Johnson, K. A., Karlawish, J., Donohue, M., Salmon, D. P., & Aisen, P. (2014). The A4 study: Stopping AD before symptoms begin? *Science Translational Medicine*, *6*(228), Article 228fs13. https://doi.org/10.1126/scitranslmed.3007941

Tombaugh, T. N., & McIntyre, N. J. (1992). The mini-mental state examination: A comprehensive review. *Journal of the American Geriatrics Society*, *40*(9), 922–935. https://doi.org/10.1111/j.1532-5415.1992.tb01992.x

Wind, A. W., Schellevis, F. G., Van Staveren, G., Scholten, R. P., Jonker, C., & Van Eijk, J. T. (1997). Limitations of the Mini-Mental State Examination in diagnosing dementia in general practice. *International Journal of Geriatric Psychiatry*, *12*(1), 101–108. https://doi.org/10.1002/(SICI)1099-1166(199701)12:1<101::AID-GPS469>3.0.CO;2-R

Received October 30, 2021
Revision received March 15, 2022
Accepted April 25, 2022 ■