Brain Imaging and Behavior
DOI 10.1007/s11682-012-9186-z

ADNI: FRIDAY HARBOR 2011 WORKSHOP SPECIAL ISSUE

Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI)

Paul K. Crane · Adam Carle · Laura E. Gibbons · Philip Insel · R. Scott Mackin · Alden Gross · Richard N. Jones · Shubhabrata Mukherjee · S. McKay Curtis · Danielle Harvey · Michael Weiner · Dan Mungas · for the Alzheimer's Disease Neuroimaging Initiative

Abstract We sought to develop and evaluate a composite memory score from the neuropsychological battery used in the Alzheimer's Disease (AD) Neuroimaging Initiative (ADNI). We used modern psychometric approaches to analyze longitudinal Rey Auditory Verbal Learning Test (RAVLT, 2 versions), AD Assessment Schedule - Cognition (ADAS-Cog, 3 versions), Mini-Mental State Examination (MMSE), and Logical Memory data to develop ADNI-Mem, a composite memory score. We compared RAVLT and ADAS-Cog versions, and compared ADNI-Mem to

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.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.ucla.edu/research/active-investigators/

Electronic supplementary material The online version of this article (doi:10.1007/s11682-012-9186-z) contains supplementary material, which is available to authorized users.

P. K. Crane (✉) · L. E. Gibbons · S. Mukherjee · S. M. Curtis
University of Washington,
Box 359780, Harborview Medical Center, 325 Ninth Avenue,
Seattle, WA 98104, USA
e-mail: pcrane@u.washington.edu

L. E. Gibbons
e-mail: gibbonsl@u.washington.edu

A. Carle
University of Cincinnati School of Medicine, Cincinnati
Children's Hospital Medical Center, and University of Cincinnati
College of Arts and Sciences,
3333 Burnet Avenue, MLC 7014,
Cincinnati, OH 45229, USA

P. Insel · R. S. Mackin · M. Weiner
Center for Imaging of Neurodegenerative Diseases (CIND),
San Francisco VA Medical Center,
4150 Clement Street,
San Francisco, CA 94121, USA

A. Gross · R. N. Jones
Department of Psychiatry, Institute for Aging Research,
Hebrew Senior Life, 1200 Center Street,
Boston, MA 02131, USA

D. Harvey
Division of Biostatistics, Department of Public Health Sciences,
University of California at Davis,
One Shields Avenue,
Davis, CA 95616, USA

D. Mungas
Department of Neurology, University of California at Davis,
4860 Y St., Suite 0100,
Sacramento, CA 95817, USA

Published online: 11 July 2012

RAVLT recall sum scores, four ADAS-Cog-derived scores, the MMSE, and the Clinical Dementia Rating Sum of Boxes. We evaluated rates of decline in normal cognition, mild cognitive impairment (MCI), and AD, ability to predict conversion from MCI to AD, strength of association with selected imaging parameters, and ability to differentiate rates of decline between participants with and without AD cerebrospinal fluid (CSF) signatures. The second version of the RAVLT was harder than the first. The ADAS-Cog versions were of similar difficulty. ADNI-Mem was slightly better at detecting change than total RAVLT recall scores. It was as good as or better than all of the other scores at predicting conversion from MCI to AD. It was associated with all our selected imaging parameters for people with MCI and AD. Participants with MCI with an AD CSF signature had somewhat more rapid decline than did those without. This paper illustrates appropriate methods for addressing the different versions of word lists, and demonstrates the additional power to be gleaned with a psychometrically sound composite memory score.

Keywords Memory · psychometrics · longitudinal analysis · cognition · hippocampus

Background

Impairments in memory are a hallmark of Alzheimer's disease (AD) and are requisite for diagnoses of the disease (McKhann et al. 1984). Assessment of memory was a crucial criterion influencing the composition of the neuropsychological battery used in the AD Neuroimaging Initiative (ADNI). The battery includes a variety of indicators of memory, including the Rey Auditory Verbal Learning Test (RAVLT) (Rey 1964), elements from the AD Assessment Scale—Cognitive Subscale (ADAS-Cog) (Mohs et al. 1997), the recall of three items from the Mini-Mental State Examination (MMSE) (Folstein et al. 1975), and recall of elements from a story from Logical Memory I of the Wechsler Memory Test-Revised (Wechsler 1987).

There are at least two reasons a memory composite score may be useful. First, summarizing all of the memory data with a single score facilitates comparisons with other variables without needing to address challenges raised by testing multiple hypotheses that would ensue if each of the memory indicators was considered separately. These other variables could be neuroimaging summaries, biomarkers, clinical diagnoses, or measures of other cognitive domains. Second, by including multiple indicators in a single score, the impact of measurement error due to idiosyncratic single items or subdomains is minimized.

Different word lists for the RAVLT and ADAS-Cog were administered at different study visits. A particular challenge that arose in these analyses was to address the two different versions of the RAVLT word lists and the three different versions of the ADAS-Cog word lists. It is important to determine whether these different versions of the RAVLT and ADAS-Cog have the same difficulty level before using total scores in longitudinal analyses. The assumption that different forms are equivalent is a strong assumption that needs to be checked (Millsap 2011). One of our goals was to compare the difficulties of the different versions of the RAVLT and ADAS-Cog used in ADNI.

Our primary goal was to develop and evaluate the validity of a psychometrically sophisticated memory composite score from the ADNI neuropsychological battery. We compared our composite memory score to a variety of other scores in a series of analyses to address the validity and performance of our composite score. First, we determined the ability of the composite to detect change over time in each diagnostic group. Second, we determined the ability to predict conversion from mild cognitive impairment (MCI) to AD. Third, we evaluated the strength of the relationship with MRI-derived parameters found in previous studies to be related to memory, including hippocampal volume, cortical thickness of the parahippocampal region, fusiform gyrus, and entorhinal cortex (Yonelinas et al. 2007; Walhovd et al. 2009; Fjell et al. 2008; Murphy et al. 2010; Van Petten et al. 2004). Finally, we compared rates of decline among people with normal cognition and with MCI who had a pattern of cerebrospinal fluid (CSF) biomarkers consistent with early AD (an "AD signature") to rates of decline among people without the AD signature.

Methods

Participants and data source

Data used in this study were obtained from the ADNI database (http://adni.loni.ucla.edu/). The ADNI was initiated in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies, and non-profit organizations. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. Michael W. Weiner, MD, VA Medical Center and University of California-San Francisco is the Principal Investigator of this initiative. This $60 million, multiyear public-private partnership involves many co-investigators from a broad range of academic institutions and private corporations. More than 800 participants, aged 55 to 90, have been recruited from across more than 50 sites

in the US and Canada. This includes approximately 200 patients diagnosed with early AD who were followed for up to 2 years. Longitudinal imaging data, including structural 1.5 Tesla MRI scans, were collected on the full sample. Neuropsychological and clinical assessments were collected at baseline, and at follow-up visits occurring at six- to twelve-month intervals. Further information about ADNI can be found in (Jack et al. 2010a) and at http://www.adni-info.org. The study was conducted after Institutional Review Board approval at each site. Written informed consent was obtained from all study participants, or their authorized representatives.

Diagnosis of amnestic MCI required patient-reported memory complaints, objective memory deficits, intact functional activities, a Clinical Dementia Rating (CDR) Scale (Morris 1993) global score of 0.5, and a MMSE (Folstein et al. 1975) score of 24 or more. Participants with AD met the National Institute of Neurological and Communicative Diseases and Stroke—Alzheimer's Disease and Related Disorders Association criteria for probable AD (McKhann et al. 1984).

Cognitive and clinical measures

Memory indicators

Considerations for compiling the ADNI neuropsychological battery included the following: 1. Coverage of the domains of interest (memory, executive functions, language, attention, and visuospatial abilities); 2. Adequate sampling of cognitive domains of interest in subjects who are normal or who have MCI or AD; 3. Can measure change over a 2–3 year period; 4. Avoid ceiling or floor effects; 5. Were efficient and met practical demands; 6. Were utilized in the AD Clinical Study (ADCS) MCI trial and worked well in that setting. Additionally, the tests are widely used in AD Centers (ADCs) that are required to collect a Uniform Data Set, to reduce the amount of testing needed for participants enrolled in ADNI from ADCs.

The RAVLT uses a 15-item list of unrelated words. This list is read to the participant, who is asked to recall aloud as many of the words as they can. The number of successfully recalled words is recorded. The list is then repeated, and the participant again asked to recall as many words as they can. This process is repeated for a total of 5 learning trials, resulting in 5 scores. Then the examiner reads a new list of 15 words to the participant (an interference word list), and the participant is asked to recall as many of these words as possible. The participant is then asked to recall the initial word list, and the number of words recalled is recorded. After thirty minutes of other testing, the participant is again asked to recall as many words from the initial list as they can. The two versions of the RAVLT include different versions of the initial and interference word lists.

The ADAS-Cog includes two different memory tasks. First is a word list learning task similar to but distinct from that of the RAVLT. The ADAS-Cog word list includes 10 unrelated words (rather than 15) that are printed on cards. The participant is asked to read them aloud (while in the RAVLT they are read to the participant) and to remember them. There are three learning trials (rather than five in the RAVLT). After five minutes (rather than 30) of unrelated testing, the participant is asked to recall as many words as possible from the list.

The second memory task included in the ADAS-Cog is a word recognition task. In this task, the participant is given 12 cards with words printed on them, and asked to read them aloud and to remember them. Then the target words along with 12 distractor words are shown to the participant, who is asked to indicate whether the word was one they were supposed to recall. Two scores are recorded: the number of target words correctly identified as being part of the list (i.e., true positives), and the number of distractor words correctly identified as not being part of the list (i.e., true negatives).

The three different versions of the ADAS-Cog include different lists of the 10 words for the list learning trial as well as different lists of the 12 words for the recognition task.

For logical memory, a brief fact-laden passage is read aloud once. The participant is asked to recall as many of the passage's 25 elements as they can, and the number of elements correctly recalled is recorded. After 30–40 minutes of other cognitive testing, the participant is again asked to recall the passage, and the number of elements correctly recalled in this delay condition is recorded.

In the MMSE, 3 words are read to the participant, who is asked to repeat them. Distractor tasks are then administered, after which the participant is asked to spontaneously recall the three words. Scores of 1 point are recorded for each item correctly recalled, and 0 for each item not correctly recalled.

Comparitor measures

We compared our composite (described below) to a variety of comparitors. The standard sum score for the five learning trials of the RAVLT was a primary comparator. Others included four versions and scores for the ADAS-Cog, including the original version (ADAS-Classic), the modified version of the ADAS-Cog that includes delayed recall (ADAS-Modified), a Rasch score developed for the original version of the ADAS-Cog (ADAS-Rasch (Wouters et al. 2008)), and a score obtained by recursive partitioning of the ADAS-Cog (ADAS-Tree (Llano et al. 2011)). Other comparitors

included the total MMSE score and the sum of boxes from the CDR.

Dementia evaluation

Conversion from normal or MCI to AD was a primary outcome for ADNI and so was tracked very closely. Complete methods for identifying dementia cases can be found in the ADNI protocol available at the ADNI web site http://www.adni-info.org.

Selected MRI-based imaging parameters

All participants had an MRI evaluation at each study visit. We identified four MRI parameters a priori as being associated with memory: hippocampal volume, thickness of the parahippocampus, thickness of the entorhinal cortex, and thickness of the fusiform gyrus. The neuroimaging methods utilized by ADNI have been described in detail previously (Jack et al., 2008) utilizing calibration techniques to maintain consistent protocols across scanners and sites. Raw dicom data of T1-weighted MP-RAGE scans acquired from 1.5 Tesla scanners at baseline visits from all participants were obtained via the ADNI database (http://www.loni.ucla.edu/ADNI/). Images were processed through FreeSurfer version 4.0.3, a software program freely available at http://surfer.nmr.mgh.harvard.edu/ to obtain measurements of hippocampal volume and cortical thickness measurements for parahippocampal, entorhinal, and fusiform gyrus regions.

CSF

A subset of participants (n=415) had baseline lumbar punctures for CSF, which was evaluated for assays of amyloid $\beta_{1\text{-}42}$ (Aβ), total tau, and phosphorylated tau_{181p} (ptau). De Meyer et al. used Aβ and ptau to classify ADNI participants as having an "AD signature" or not (De Meyer et al. 2010), and provided us with the classes for these analyses.

Psychometric analyses of baseline data

Our initial modeling of memory focused on baseline data to determine whether a single factor model would be appropriate or whether a more complicated model would be necessary.

We used Mplus statistical software for all models (Muthén and Muthén 2006). Mplus facilitates very flexible modeling but allows a maximum of 10 categories per categorical indicator. We re-coded memory indicators to have a maximum of 10 categories. We developed a re-coding algorithm based on preserving variability at the extremes of the distribution at the expense of variability in the middle range of the distribution. Specific re-coding we used is shown in Table S1.

We compared a single factor model to a bi-factor model that included additional terms to capture covariance not due to the underlying factor defined by all of the indicators (McDonald 1999; Reise et al. 2007). Our initial task was then to identify one or more specific candidate bi-factor models to compare with the single factor model. We considered two approaches: one accounting for theoretical considerations regarding memory subtypes assessed by each of the indicators, and the other accounting for methods effects.

For the first approach, before we looked at data we (P.K.C., A.C., and D.M.) assigned memory indicators from the ADNI data set into categories based on the memory subtype it assessed ("content" models). Specific subtypes we considered were list learning and paragraph recall. For the second approach, we considered whether the same stimulus was being assessed several times ("methods" models). For example, for the ADAS-Cog, there were three word list learning trials and a recall trial of the same list of words, while the recognition task was of a different list of words but had both true and false positives. We thus modeled a secondary methods factor for the first four indicators which would capture the facility people had with those specific words beyond their overall memory ability, and a secondary residual correlation between the true and false positives for the recognition task, which captures additional covariation between those indicators beyond their relationship with overall memory.

We compared these candidate secondary domain structures on the basis of published desirable thresholds for the fit statistics (Reeve et al. 2007). We specifically focused on the confirmatory fit index (CFI), where values >0.95 are consistent with excellent fit; on the Tucker-Lewis Index (TLI), where values >0.95 are consistent with excellent fit; and on the root mean squared error of approximation (RMSEA), where values <0.08 are consistent with adequate fit and values <0.05 are consistent with excellent fit. Based on these analyses, the bi-factor model with methods effects was far superior to the content bi-factor model, so we only considered the methods effects model in subsequent analyses.

Finally we compared the single factor and the methods bi-factor models. We noted the fit indices for these two models, though fit statistics were not deciding criteria. Much more important for our purposes was the correlation between memory factor scores from the two models, and the scatter plot showing the relationship between these scores. We also compared the loadings for each indicator on the overall memory factor, with and without the secondary domain structure.

Mplus code for all of these analyses is available on request from the first author.

Psychometric analyses of longitudinal data

The task of modeling the longitudinal memory data was complicated by the multiple forms of the ADAS-Cog word lists and the RAVLT word list. Furthermore, Logical Memory I was only assessed at annual visits. The only indicators consistently present across visits were the three word recall items from the MMSE. Technically these three dichotomous indicators (i.e., correct / incorrect) could be used to anchor the scales across time points (Steven P. Reise et al. 1993), but we were concerned that this anchoring would be too sparse for firm conclusions to be drawn. Because of the multiple versions of the RAVLT and the ADAS-Cog administered at different ADNI study visits, we needed to use longitudinal data to establish our final composite scores, since we could not assume that the different versions were of the same difficulty.

Based on results from initial cross-sectional modeling described above, we limited ourselves to single factor models. We divided the data set into two parts: first, the annual visits (baseline, month 12, and month 24), and second, the other visits (month 6, 18, and 36). Logical Memory I and II were assessed at each of the visits in the first half of the data set, so those much richer indicators were used as anchors alongside the three dichotomous MMSE indicators. Furthermore, at each of those visits, only the first version of the RAVLT was assessed, so it could also act as an anchor. The only thing that varied at those visits was thus the three different versions of the ADAS-Cog. We fit a longitudinal model using all available data for the annual visits of the first half of the data set. We identified the scale by specifying the variance of the general factor to be 1 at the baseline visit, when its mean was 0. We allowed the mean and the variance of the general factor to vary at other time points, and the general factors were freely correlated with each other. We freely estimated the loadings on the general factor, but constrained those loadings from the same indicators to be the same across time points. For example, for the first MMSE item, we freely estimated the loading on the overall memory factor at each time point, but constrained that loading to be the same at baseline, month 12, and month 24.

We captured point estimates for the loadings and thresholds for the three MMSE items, Logical Memory I and II, and the three versions of the ADAS-Cog from the first half of the data set. We then turned our attention to the second half of the data set that included data from study visits at months 6, 18, and 36. The second version of the RAVLT word list was used at each of these study visits. We used the MMSE items, the ADAS-Cog version 2 (month 6), version 1 (month 18), and version 3 (month 36), and Logical Memory (month 36) as anchors to estimate item parameters for the second version of the RAVLT. The longitudinal modeling strategy was similar to that described for the first half of the data. Because we were fixing item loadings and thresholds for the anchor items, the scale was still anchored to the mean of 0 and variance of 1 at the baseline visit, we freely estimated the means and variances at each of the study visits included in this second half of the data. Script files for these analyses are available on request.

We extracted factor scores for each participant at each study visit (named ADNI-Mem in the ADNI data set). We compared item parameters (factor loadings and category thresholds) across the three different versions of the ADAS-Cog and the two different versions of the RAVLT.

Mplus code for all of these analyses is available on request from the first author.

Comparisons of scores

We performed several analyses to compare our memory composite to other scores.

Rates of change We examined the sensitivity of each measure to change over time in each of the three diagnostic groups using z-statistics based on the coefficients and standard errors for time from mixed models for the cognitive outcomes using random intercepts and slopes and an unstructured covariance matrix, controlling for age, education, sex and presence of one or more APOE ε4 alleles. We used the coefficients for year and the adjusted residual standard deviation from these models to determine sample sizes needed per group to detect a 25 % reduction in the rate of decline in 12 months for a two-arm trial, with 80 % power and alpha = 0.05, assuming a two-sided test.

Time to conversion for people with MCI We compared the strength of association between cognition and risk of developing dementia, using accelerated failure time models of time to AD, with a Weibull distribution, controlling for age, education, sex, and presence of one or more APOE ε4 alleles. We performed two sets of analyses. First, we evaluated baseline cognitive scores. Second, we performed a lagged analysis to compare the strength of association between cognitive variability at each visit and risk of developing dementia at the subsequent study visit.

Strength of association with MRI parameters We determined the strength of association between cognitive scores and

Table 1 Demographic, clinical, CSF and MRI data by baseline diagnosis (n=803 with complete cognitive data)

	Normal cognition	Mild cognitive impairment (MCI)	Alzheimer's disease (AD)
Sample size, N			
Baseline	225	394	184
6 months	215	371	171
12 months	202	351	149
18 months	0	316	0
24 months	193	289	115
36 months	164	209	0
Demographics			
Female	48 %	35 %	49 %
Age (years), mean (SD)	76.0 (5.0)	74.9 (7.5)	75.5 (7.4)
Education (years), mean (SD)	16.0 (2.8)	15.7 (3.0)	14.7 (3.1)
Any *APOE* ε4 alleles	26 %	53 %	65 %
Baseline clinical data: mean (SD)			
Memory			
ADNI-Mem	1.0 (0.5)	−0.1 (0.6)	−0.8 (0.5)
RAVLT Trials 1–5 sum	43.3 (9.1)	30.8 (9.0)	23.3 (7.5)
Global Cognition			
ADAS-Classic (70 pts)	6.2 (2.9)	11.5 (4.4)	18.5 (6.3)
ADAS-Total (85 pts)	9.4 (4.2)	18.6 (6.3)	28.8 (7.7)
ADAS-Rasch	4.8 (3.5)	11.8 (5.5)	19.5 (7.4)
ADAS-Tree	7.9 (3.5)	15.9 (5.1)	24.2 (5.6)
MMSE	29.1 (1.0)	27.0 (1.8)	23.4 (2.0)
Clinical Rating			
CDR-SB	0.0 (0.1)	1.6 (0.9)	4.3 (1.7)
Baseline CSF data			
CSF data, N	112	193	97
de Meyer AD cluster, %	35 %	73 %	90 %
Baseline MRI data: mean (SD)			
Hippocampus volume, cm^3	6.7 (0.8)	5.8 (1.0)	5.2 (1.0)
Entorhinal thickness, mm	3.5 (0.3)	3.1 (0.5)	2.7 (0.5)
Fusiform thickness, mm	2.6 (0.2)	2.5 (0.2)	2.3 (0.3)
Parahippocampal thickness, mm	2.6 (0.3)	2.4 (0.3)	2.3 (0.3)
Complete data for MRI, N	185	305	118

Abbreviations: ADAS: Alzheimer's Disease Assessment Schedule; CDR-SB: Clinical Dementia Rating Scale—Sum of Boxes; CSF: Cerebrospinal fluid; mm: millimeter; MMSE: Mini-Mental State Examination; MRI: Magnetic resonance imaging; RAVLT: Rey Auditory Verbal Learning Test; SD: Standard deviation

selected MRI values from baseline in each of the diagnostic groups using linear regression models predicting the cognitive outcome, adjusting for total intracranial volume, age, education, sex, and presence of one or more APOE ε4 alleles.

Ability to differentiate trajectories of participants with CSF AD signatures among people with normal cognition and with MCI We used mixed effects models to determine the ability of each cognitive measure to differentiate the cognitive trajectories of participants with an AD profile of CSF biomarkers compared to people without that profile. Our rationale for limiting these analyses to participants with normal cognition and with MCI was that people ultimately destined to develop AD should have greater rates of decline in cognition in general and memory in particular than people not destined to develop AD, but that the AD CSF profile might not have a relationship with subsequent trajectories of cognition among people with established AD (Jack et al. 2010b). Analyses were conducted within each diagnostic group with random intercepts and slopes and an unstructured covariance matrix, controlling for age, education, sex, and presence of one or more APOE ε4 alleles.

Results

Characteristics of participants

Of the 819 ADNI participants eligible at baseline, 803 had complete data for our cognitive outcomes at one or

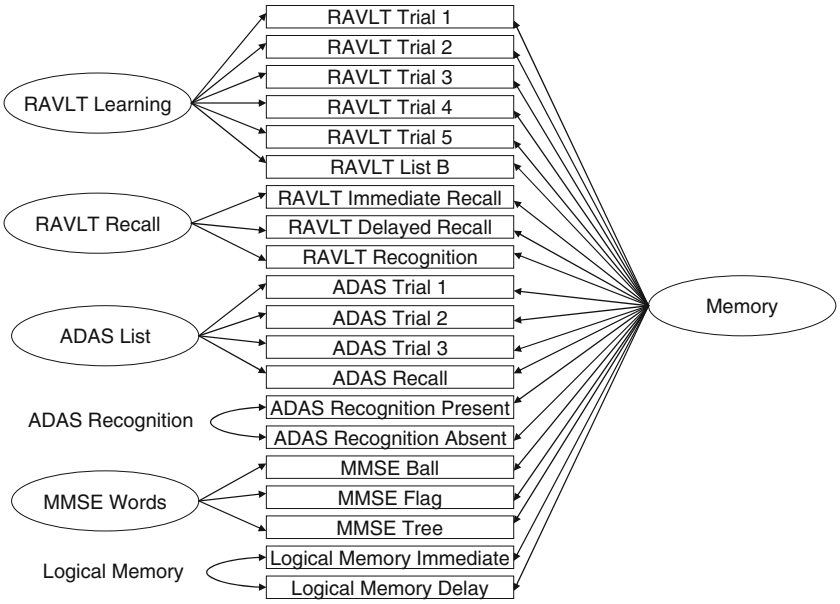

Fig. 1 Bi-factor model path diagram for baseline data. RAVLT=Rey Auditory Verbal Learning Test. ADAS=Alzheimer's Disease Assessment Schedule. MMSE=Mini-Mental State Examination. Covariation across all the indicators is modeled with loadings on the primary "Memory" factor shown to the right. Shared covariation beyond that shared with all of the items is shown in secondary factors (for three or more indicators) and residual correlations (for two indicators, shown as two-headed curved arrows) to the left. For example, shared covariation for the 6 word list learning trials for the RAVLT (five with list A, one with list B) beyond that shared with all the other indicators is modeled with the "RAVLT Learning" factor. We specified a unit variance for each of the factors, and they were mutually uncorrelated with each other

more study visits. Of these, 225 had normal cognitive functioning, 394 had mild cognitive impairment (MCI), and 184 had AD. Demographic, clinical, CSF, and imaging data for these individuals are shown in Table 1.

Cross-sectional analyses of memory indicators

We compared candidate bi-factor models as described in the Methods section. Our best-fitting candidate model had secondary domains for methods effects, and split the RAVLT into a learning factor (including the interference list) and a recall factor. The path diagram for the selected bi-factor model is shown in Fig. 1. Loadings for the bi-factor model are shown in Table 2. The first column of data shows standardized loadings for the overall "Memory" factor. The second column of data shows loadings for the relevant subdomain. We shaded the rows to highlight membership of particular memory indicators in particular subdomains. Two pairs of items had residual correlations rather than underlying factors; we show the residual correlation in one row of the table and place one or two asterisks in the corresponding row of the partner indicator. All of the standardized factor loadings on the overall "Memory" factor were well over 0.30, McDonald's threshold for salience (McDonald 1999), suggesting that all of the items—including the three dichotomous MMSE words—are salient indicators of overall memory. For each indicator, loadings on the overall "Memory" factor were higher than the loading on the method subdomain factor. Several of the loadings on the method subdomain factors were below the 0.30 threshold for salience. There was a negative correlation between the true and false positive indicators for the ADAS-Cog recognition task. The factor loadings for these two items indicate that both true hits and true misses are salient indicators of overall memory, and that they have a negative residual correlation, meaning that beyond their overall relationship with memory they have a negative relationship with each other. We suspect this reflects the effects of strategies for guessing. If a respondent is not sure whether a candidate word was truly presented and guesses, and has a strategy of guessing "present," then the number of true hits will be higher and the number of true misses will be lower; conversely, if a respondent has a strategy of guessing "absent," then the number of true hits will be higher and the number of true misses will be lower. Taken together, these strategies for guessing result in a negative residual correlation—the parts of these scores not reflecting overall memory are negatively related to each other.

We compared the bi-factor model described above to the single factor model that assumed no residual relationships. The bi-factor model fit the data better than a single factor model. For the bi-factor score, the CFI was 0.97, the TLI was 0.99, and the RMSEA was 0.086. For

Table 2 Factor loadings for the primary and secondary factors for the bi-factor model from baseline

	Loading on primary factor	Loading on secondary factor or residual correlation	Name of secondary domain
RAVLT trial 1	0.55	0.49	RAVLT Learning
RAVLT trial 2	0.75	0.48	
RAVLT trial 3	0.84	0.35	
RAVLT trial 4	0.89	0.27	
RAVLT trial 5	0.91	0.20	
RAVLT interference trial	0.58	0.23	
RAVLT immediate recall	0.85	0.28	RAVLT Recall
RAVLT 30 minute delay	0.71	0.41	
RAVLT recognition	0.88	0.24	
ADAS trial 1	0.76	0.38	ADAS List
ADAS trial 2	0.81	0.47	
ADAS trial 3	0.80	0.38	
ADAS recall	0.88	0.17	
ADAS recognition present	0.41	-0.33	ADAS Recognition
ADAS recognition absent	0.50	*	
MMSE ball	0.60	0.48	MMSE Words
MMSE flag	0.67	0.52	
MMSE tree	0.63	0.52	
Logical Memory Immediate	0.78	0.24	Logical Memory
Logical Memory Delay	0.80	**	

* and ** indicate residual correlations.

Abbreviations: ADAS: Alzheimer's Disease Assessment Schedule; MMSE: Mini-Mental State Examination; RAVLT: Rey Auditory Verbal Learning Test

the single factor model, the CFI was 0.89, the TLI was 0.97, and the RMSEA was 0.179.

Category thresholds are determined from the proportions of people responding in each category, and threshold values for all indicators are identical for the single and bi-factor models; the only difference between the models was to be found in the factor loadings. We show a comparison of the factor loadings in Table 3. As expected, most loadings on the general factor were somewhat attenuated in the bi-factor model compared to the single factor model, since some of the covariation assumed

Table 3 Factor loadings on the general (overall memory) factor for the single factor and bi-factor models

Indicator	Loading for single factor model	Loading for bi-factor model	Absolute difference	Difference as percent of single factor loading
RAVLT trial 1	0.62	0.55	0.07	11 %
RAVLT trial 2	0.82	0.75	0.06	8 %
RAVLT trial 3	0.88	0.84	0.04	4 %
RAVLT trial 4	0.92	0.89	0.02	2 %
RAVLT trial 5	0.92	0.91	0.01	1 %
RAVLT interference trial	0.60	0.58	0.02	3 %
RAVLT immediate recall	0.87	0.85	0.01	2 %
RAVLT 30 minute delay	0.71	0.71	0.00	0 %
RAVLT recognition	0.90	0.88	0.01	2 %
ADAS trial 1	0.79	0.76	0.03	4 %
ADAS trial 2	0.86	0.81	0.05	6 %
ADAS trial 3	0.84	0.80	0.03	4 %
ADAS recall	0.87	0.88	−0.01	−1 %
ADAS recognition present	0.39	0.41	−0.03	−6 %
ADAS recognition absent	0.48	0.50	−0.03	−5 %
MMSE ball	0.60	0.60	0.01	1 %
MMSE flag	0.67	0.67	0.00	0 %
MMSE tree	0.63	0.63	0.01	1 %
Logical Memory Immediate	0.83	0.78	0.06	7 %
Logical Memory Delay	0.85	0.80	0.05	6 %

Abbreviations: ADAS: Alzheimer's Disease Assessment Schedule. MMSE: Mini-Mental State Examination. RAVLT: Rey Auditory Verbal Learning Test

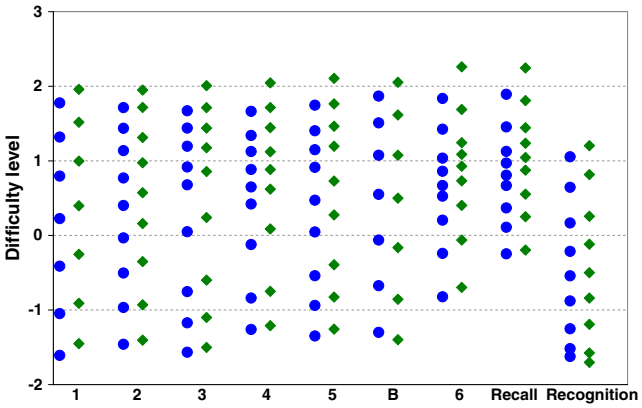

Fig. 2 Difficulty levels for the elements of the two versions of the Rey Auditory Verbal Learning Test. The five learning trials are indicated by the numbers 1 through 5; the interference trial by the letter B, the first recall trial by the number 6; delayed recall by "Recall", and the recognition task by "Recognition". Version A difficulty thresholds are denoted with blue circles, while version B difficulty thresholds are denoted with green diamonds. In this plot, the difficulty levels are plotted on the y axis in z-statistic units; higher numbers indicate higher memory ability / higher item difficulty. Considering the two versions of learning trial 1, version A is easier for each threshold. At an overall memory ability level of −0.5, for example, higher proportions of people will be above the first threshold for version A, and lower proportions of people above that same threshold for version B. At every threshold the green diamonds are higher than the blue dots. For the second through 5th learning trials, this difference is dramatic at the top end, as the top threshold on version A is only as difficult as the 2nd to highest threshold on version B

to be related to the general factor in the single factor model was modeled in secondary domains and residual correlations in the bi-factor model. The largest absolute difference was for Trial 1 of the RAVLT, which had loadings of 0.62 in the single factor model and 0.55 in the bi-factor model, a difference of

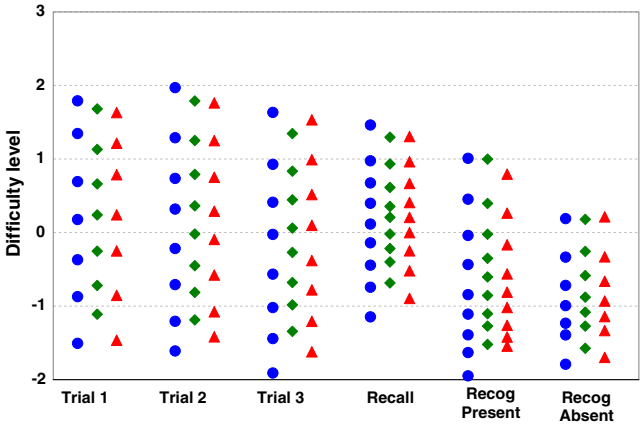

Fig. 3 Difficulty levels for the elements of the three versions of the Alzheimer's Disease Assessment Scale – Cognitive Subscale. Recog=Recognition. Version A threshold difficulty levels are depicted with blue circles, Version B with green diamonds, and Version C with red triangles. In this plot, the difficulty levels are plotted on the y axis in z-statistic units; higher numbers indicate higher memory ability / higher item difficulty Version A has greater spread than Version B and to a lesser extent than version C, meaning it will have slightly smaller ceiling and floor effects. Unlike the Rey, no version appears to be consistently easier or harder than the other versions

Table 4 Coefficients for time, in mixed models for cognition controlling for age, education, gender and presence of one or more APOE ε4 alleles. Bold font indicates $p<0.05$. Sample size needed per group to detect a 25 % decrease over 12 months, with 80 % power and alpha = 0.05, two-sided. ADAS and CDR-SB scores reversed so that higher scores represent better cognition for all clinical measures

Clinical outcome	Time			Sample size per group*		
	NC	MCI	AD	NC	MCI	AD
Memory						
ADNI-Mem	**3.02**	**−9.43**	**−11.59**	28,512	2167	568
RAVLT Trials 1–5 sum	−0.53	**−8.67**	**−10.13**	1,333,396	4292	804
Global Cognition						
ADAS-Classic (70 pts)	**3.20**	**−10.78**	**−12.25**	37,971	1651	242
ADAS-Total (85 pts)	1.76	**−12.20**	**−12.92**	105,895	1200	206
ADAS-Rasch	**3.10**	**−10.51**	**−11.28**	41,295	1692	346
ADAS-Tree	0.73	**−13.67**	**−14.05**	573,996	981	214
MMSE	−0.36	**−11.27**	**−9.95**	3,494782	1628	393
Clinical Rating						
CDR-SB	**−4.06**	**−14.57**	**−13.64**	4,315**	456	223

* Sample size calculations are based on the coefficient for year and the adjusted residual standard deviation from the full model. As such, they may not directly correspond to the z-statistics.

** Note that the CDR-SB has very skewed data among participants with NC, so assumptions made by the mixed effects models probably do not hold. Thus this value is likely inaccurate but included here for comparison purposes.

Abbreviations: AD: Alzheimer's disease. ADAS: Alzheimer's Disease Assessment Schedule. ADNI-Mem: Alzheimer's Disease Neuroimaging Initiative Memory Score. CDR-SB: Clinical Dementia Rating Scale—Sum of Boxes. MCI: Mild cognitive impairment. MMSE: Mini-Mental State Examination. NC: Normal cognition. RAVLT: Rey Auditory Verbal Learning Test

Table 5 Time ratios (TR), with 95 % confidence intervals (CI), for predicting conversion to dementia, controlling for age, education, gender and presence of one or more APOE ε4 alleles. ADAS and CDR-SB scores reversed so that higher scores represent better cognition for all clinical measures*

Clinical predictor	Previous visit Time Ratio (95 % CI)	Baseline Time Ratio (95 % CI)
Memory		
ADNI-Mem	1.50 (1.32, 1.70)	1.53 (1.34, 1.74)
RAVLT Trials 1–5 sum	1.40 (1.24, 1.57)	1.37 (1.22, 1.54)
Global cognition		
ADAS-Classic (70 pts)	1.36 (1.23, 1.51)	1.42 (1.25, 1.61)
ADAS-Total (85 pts)	1.45 (1.29, 1.63)	1.50 (1.31, 1.71)
ADAS-Rasch	1.25 (1.14, 1.38)	1.37 (1.22, 1.55)
ADAS-Tree	1.48 (1.31, 1.67)	1.54 (1.35, 1.77)
MMSE	1.23 (1.13, 1.34)	1.36 (1.15, 1.61)
Clinical Rating		
CDR-SB	1.49 (1.29, 1.71)	1.46 (1.24, 1.73)

* We used an accelerated hazard model with a Weibull distribution to account for interval censoring in the data. Adjusted time ratios greater than one indicate a longer time until progression to dementia.

Abbreviations: ADAS: Alzheimer's Disease Assessment Schedule. ADNI-Mem: Alzheimer's Disease Neuroimaging Initiative Memory Score. CDR-SB: Clinical Dementia Rating Scale—Sum of Boxes. MMSE: Mini-Mental State Examination. RAVLT: Rey Auditory Verbal Learning Test.

0.07, or 11 % of the single factor loading. None of the other indicators had differences as large as 10 %. As expected, when ignoring the negative residual correlation between the recognition tasks for the ADAS-Cog, the loadings on the primary factor were somewhat smaller. Differences in loadings for those two indicators were small between the single factor and the bi-factor model, and loadings on the overall factor were still over the 0.30 threshold for salience.

The overall correlation between single-factor and bi-factor scores for memory at the baseline exam was 0.99. The correlation for participants with AD was 0.98; for participants with MCI it was 0.99; for participants with normal cognition it was 0.98. A scatter plot did not suggest any systematic differences from the diagonal (Figure S1).

These results suggested that a single-factor model was appropriate for our purposes, as there was negligible difference between single-factor and bi-factor scores.

Version effects for the RAVLT and the ADAS-Cog

The loadings for each of the indicators from the two versions of the RAVLT were very similar (Table S2); as a proportion, they ranged from 5 % smaller to 3 % larger between the two versions. The difficulty levels for the category thresholds, however, displayed important differences between the two versions, as shown in Fig. 2. The values for the thresholds

Table 6 Coefficients for MRI thickness measures from regression models for the cognitive measure controlling for age, education, gender, presence of one or more APOE ε4 alleles, and intracranial volume. Bolded coefficients indicate p-values < 0.05. ADAS and CDR-SB scores reversed so that higher scores represent better cognition for all clinical measures

Clinical outcome	Hippocampal volume	Parahippocampal thickness	Entorhinal thickness	Fusiform thickness
Normal Cognition				
Memory				
ADNI-Mem	0.27	0.55	−0.91	**−1.98**
RAVLT Trials 1–5 sum	0.65	0.58	−0.91	**−2.02**
Global Cognition				
ADAS-Classic (70 pts)	1.03	1.30	0.27	−0.17
ADAS-Total (85 pts)	0.36	1.23	−0.12	−0.90
ADAS-Rasch	0.16	1.33	0.28	−0.52
ADAS-Tree	−0.26	0.76	−0.39	−1.44
MMSE	−0.16	1.30	0.28	0.10
Clinical Rating				
CDR-SB	−0.32	0.51	1.69	1.21
Mild Cognitive Impairment				
Memory				
ADNI-Mem	**6.72**	**3.02**	**7.59**	**4.75**
RAVLT Trials 1–5 sum	**4.23**	**2.51**	**4.95**	**3.95**
Global Cognition				
ADAS-Classic (70 pts)	**6.32**	1.63	**7.68**	**4.46**
ADAS-Total (85 pts)	**7.16**	1.84	**8.34**	**4.72**
ADAS-Rasch	**5.63**	1.76	**5.47**	**4.90**
ADAS-Tree	**7.79**	**2.10**	**8.21**	**4.59**
MMSE	**2.86**	1.10	**2.80**	**3.27**
Clinical Rating				
CDR-SB	**2.91**	**2.84**	**3.38**	**2.80**
Alzheimer's Disease				
Memory				
ADNI-Mem	**3.31**	**2.09**	**3.33**	**2.26**
RAVLT Trials 1–5 sum	**2.04**	1.11	**2.02**	1.46
Global Cognition				
ADAS-Classic (70 pts)	**2.50**	1.73	**4.60**	3.20
ADAS-Total (85 pts)	**2.67**	**2.13**	**4.67**	**3.38**
ADAS-Rasch	1.59	0.86	**3.17**	**3.05**
ADAS-Tree	**3.28**	**2.48**	**4.52**	**3.13**
MMSE	**3.01**	0.98	**3.38**	**2.99**
Clinical Rating				
CDR-SB	**2.76**	1.35	1.74	0.97

Abbreviations: ADAS: Alzheimer's Disease Assessment Schedule. ADNI-Mem: Alzheimer's Disease Neuroimaging Initiative Memory Score. CDR-SB: Clinical Dementia Rating Scale—Sum of Boxes. MMSE: Mini-Mental State Examination. RAVLT: Rey Auditory Verbal Learning Test

between item categories are plotted on the Y axis. Version 1 thresholds are shown in blue circles, while version 2 thresholds are shown in green diamonds. For all of the trials with the exception of List B (the distractor list), the Version 2 list is more difficult (has higher thresholds) than the Version 1 list. As expected, recall is more difficult than recognition (see two right-most sets of thresholds). These differences in difficulty thresholds mean RAVLT total scores for any person with high memory ability levels would be expected to differ by 5 or 6 points entirely as a function of which version of the test was used. For people with lower memory ability levels, expected differences in RAVLT total scores are smaller, but the expected difference would still be 2 or 3 points entirely as a function of which version of the test was used.

The ADAS-Cog versions were more similar to each other, at least in terms of category thresholds (see Fig. 3). Version 1 had a greater spread of thresholds than Version 2 and to a lesser extent than Version 3, which means that it should be somewhat better able to differentiate among people at the extremes of memory ability with fewer ceiling or floor scores. The loadings for the learning trials and recall of the three versions of the ADAS-Cog list learning task were very similar

Table 7 Z-scores for the slope and intercept of CSF-based AD signature group from mixed models for change in the cognitive outcomes, controlling for age, education, sex and presence of one or more APOE ε4 alleles. Bolded coefficients indicate p-values < 0.05. ADAS and CDR-SB scores reversed so that higher scores represent better cognition for all clinical measures

Clinical outcome	NC		MCI	
	Intercept	Slope	Intercept	Slope
Memory				
ADNI-Mem	0.04	0.00	**−3.20**	**−5.19**
RAVLT Trials 1–5 sum	−0.07	0.54	**−2.84**	**−3.60**
Global Cognition				
ADAS-Classic (70 pts)	−0.65	**−1.96**	**−1.97**	**−4.39**
ADAS-Total (85 pts)	−0.46	**−2.08**	**−3.09**	**−4.64**
ADAS-Rasch	−0.23	−1.71	**−2.18**	**−4.40**
ADAS-Tree	−0.26	−1.68	**−3.62**	**−4.74**
MMSE	0.47	−0.53	−1.86	**−4.48**
Clinical Rating				
CDR-SB	0.83	**−2.55**	−1.80	**−5.14**

Abbreviations: AD: Alzheimer's disease. ADNI-Mem: Alzheimer's Disease Neuroimaging Initiative Memory Score. ADAS: Alzheimer's Disease Assessment Schedule. CDR-SB: Clinical Dementia Rating Scale—Sum of Boxes. CSF: Cerebrospinal fluid. MCI: Mild cognitive impairment. MMSE: Mini-Mental State Examination. NC: Normal cognition. RAVLT: Rey Auditory Verbal Learning Test.

to each other, with differences ranging from 4 percent lower to 2 percent higher (Table S3). The recognition present and recognition absent tasks had somewhat dissimilar loadings. In no case were these strong indicators of overall memory (standardized loadings ranged from 0.43 to 0.56, roughly half the magnitude of loadings for the list learning indicators). The largest overall difference in loading between versions was 0.13 for recognition correct between Version A and Version C, which in terms of percentage was a 30 % difference in loadings.

Comparison of the ADNI-Mem to other measures

Table 4 shows the standardized coefficients for change over time for our ADNI-Mem composite score and for the comparison measures. The table highlights the two tests of memory (ADNI-Mem and the RAVLT) in the top section, and proceeds to address tests of global cognition (several scores derived from the ADAS-Cog and the MMSE) and a global clinical measure (the CDR sum of boxes). There is not much change that occurs over the course of two years for ADNI participants with normal cognition. This is reflected in the small standardized coefficients for all of the measures. Indeed, on average, ADNI-Mem and two of the global ADAS-Cog scores indicate very modest improvement in cognition over two years (positive coefficients). Among people with MCI, ADNI-Mem performed somewhat better than the RAVLT sum score, and nearly as well as the global ADAS-Cog scores or the clinical CDR sum of boxes. Among people with AD, all of the scores are able to detect robust changes over time, and ADNI-Mem performed somewhat better than the RAVLT total score.

Table 5 shows results for the ability of the scores to predict conversion to dementia. Results appeared similar for all of the scores, though time ratios (the equivalent of hazard ratios had we used Cox models) for ADNI-Mem were either the best or second best among all of the measures assessed.

Table 6 shows results for the cross-sectional association of each score with four neuroimaging parameters from MRI. Findings among people with normal cognition are difficult to understand, as there is a statistically significant inverse relationship between fusiform thickness and our ADNI-Mem composite score. This inverse relationship was also present for the RAVLT total score. For people with MCI, there were strong associations in the expected direction between ADNI-Mem and all four neuroimaging markers, suggesting that poorer memory performance was associated with smaller hippocampal volumes and with thinner cortex in the parahippocampal, fusiform, and entorhinal regions. Further, in each case the strength of association for these imaging findings was somewhat stronger than that for the total RAVLT score, and comparable to that of the various versions of the ADAS-Cog. Among people with AD, there was again a strong association between ADNI-Mem and each of the imaging parameters, and the strength of this association was somewhat stronger in each case than that for the RAVLT total score.

Table 7 shows results for the differences in intercept and rates of decline associated with having an AD CSF signature for people with normal cognition and MCI. Among people with normal cognition, there was little difference in trajectories associated with having the AD CSF signature, though there were differences in trajectories for the modified ADAS-Cog and the CDR sum of boxes in the hypothesized direction (i.e., people with the AD CSF signature had faster rates of decline). Among people with MCI, all of the measures considered suggested faster rates of decline among people with the AD CSF signature. This difference was largest for ADNI-Mem.

Discussion

In this paper we present methods we used to derive a memory composite from the neuropsychological battery administered in ADNI. We found a single factor model to be quite acceptable for the memory indicators from this battery. Our composite addresses an under-appreciated

challenge in these data, which is that the study administered three different versions of the ADAS-Cog word lists and two different versions of the RAVLT word lists. We found that the ADAS-Cog item thresholds were similar across versions, though the relative importance of the recognition tasks varied somewhat. For the RAVLT, on the other hand, we found an important difference in difficulty levels, as the second version of RAVLT was systematically more difficult than the first version. Failing to account for these differences in difficulty levels could result in strange results if standard sum scores are used. Our memory composite performed well in comparison to other cognitive measures. It was able to detect change over time well among people with MCI and AD. It was a strong predictor of conversion from MCI to AD. It was strongly associated with a priori specified neuroimaging parameters selected on the basis of their known association with memory performance. It was able to detect differences in changes over time for people with MCI who had CSF biomarkers suggesting an AD signature.

These results suggest that the two RAVLT word lists used in ADNI are not equivalent to each other (list 2 is systematically harder than list 1). If standard total scores are used, this may result in artifactual saw-tooth patterns in plots of performance over time, since people with no change in actual memory performance would be expected to have higher scores / lower scores / higher scores / lower scores at alternating visits. Because of the design of the study, participants with AD did not have an 18-month study visit, so their four observations would have the pattern higher scores / lower scores / higher scores / higher scores. The scoring approach adopted for ADNI-Mem accounts for the different difficulty levels of the two versions of the RAVLT. We did not account for different versions of the RAVLT when using changes in the RAVLT in analyses; we are not familiar with traditional methods for doing so, and to our knowledge different version effects have not been considered in publications that have analyzed ADNI RAVLT data.

The three versions of the ADAS-Cog were much more similar to each other than were the two versions of the RAVLT to each other. Nevertheless, there were differences in the relative importance of the recognition tasks across the different versions of the ADAS-Cog. Attention could be paid to the relative importance of these recognition tasks in the different versions of the ADAS-Cog, especially if the scoring to be applied to these versions does not account for this.

The ADNI-Mem composite score has several desirable features. It appears to have good validity, as it performed as well or better than the RAVLT in each of the analyses performed. Unlike the standard sum scores used for the RAVLT, however, ADNI-Mem accounts for the different versions of the RAVLT and the ADAS-Cog. ADNI-Mem also includes additional information from logical memory and from the MMSE, incorporating all of the memory-related information available from the neuropsychological battery administered in ADNI. Basing inferences on a multiple indicator composite rather than single measures conserves statistical power by reducing the number of potential comparisons, and may reduce measurement error. It uses a sophisticated modern psychometric approach that is based entirely on inter-relationships among items rather than external criteria such as those used in the recursive partitioning approach that generated the ADAS-Tree scores. The modern psychometric approach used to generate the ADNI-Mem scores has linear scaling properties that are appropriate for tracking changes over time (Crane et al. 2008; Mungas and Reed 2000).

The rationale for using the ADNI-Mem score in analyses of ADNI data is thus multifaceted. From a theory perspective, it has many desirable properties. These include incorporating all memory indicators, thus maximizing measurement precision of the memory level underlying responses to memory items; it has linear scaling properties that are especially important in longitudinal analyses; and it accounts for version effects in the RAVLT and ADAS-Cog. From a data-driven perspective, it also has desirable properties: it appears to be at least as valid as its constituent parts, and did well in predicting people who would progress from MCI to AD and in detecting changes over time. We have submitted our ADNI-Mem scores to the ADNI data base and recommend their use by any researcher using the ADNI data set who has substantive questions about memory. Specifically, the ADNI-Mem scores may be particularly useful for imaging researchers who wish to compare image processing and analysis techniques in terms of the strength of associations between imaging and memory.

Limitations should be considered in interpreting our results. We were limited by the battery of tests administered by ADNI. We suspect—but cannot confirm—that similar findings would have been obtained had other tests been used. Although the ADNI battery is fairly rich in its assessment of memory, the advantages of a composite score approach would presumably be even more apparent if even more tests were available. We did not compare the ADNI neuropsychological battery to any other battery of tests, and cannot comment on whether it may be superior to other batteries used clinically or in other research studies. The ADNI data set includes rich neuroimaging results available from study participants, making it an ideal setting for our analyses comparing various scores to imaging findings. We selected four specific measures a priori. Had we selected different measures we could have found different findings. Similarly, there are a variety of ways of estimating hippocampal volume. We relied on one particular technique. Only a subset of the ADNI sample had CSF measures. Our findings would have been more robust had our sample sizes for the CSF analyses been larger.

In conclusion, this paper outlines the methods for developing the ADNI-Mem composite measure of memory for the ADNI study, and compares it to several other cognitive tests. We also found that the two versions of the RAVLT are of very different difficulty levels, a fact that is accounted for in the composite ADNI-Mem scores. The ADNI-Mem scores should be used when a single indicator of memory performance is desired. We have supplied these scores so that they are available in the ADNI data set.

Acknowledgment Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. 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 Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory of Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation. Data management and the specific analyses reported here wer\e supported by NIH grant R01 AG029672 (Paul Crane, PI), P50 AG05136 (Murray Raskind, PI), and R13 AG030995 (Dan Mungas, PI).

References

Crane, P. K., Narasimhalu, K., Gibbons, L. E., Mungas, D. M., Haneuse, S., Larson, E. B., et al. (2008). Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. *Journal of Clinical Epidemiology*, *61*(10), 1018–1027 e1019.

De Meyer, G., Shapiro, F., Vanderstichele, H., Vanmechelen, E., Engelborghs, S., De Deyn, P. P., et al. (2010). Diagnosis-independent Alzheimer disease biomarker signature in cognitively normal elderly people. *Archives of Neurology, 67*(8), 949–956. doi:10.1001/archneurol.2010.179.

Fjell, A. M., Walhovd, K. B., Amlien, I., Bjornerud, A., Reinvang, I., Gjerstad, L., et al. (2008). Morphometric changes in the episodic memory network and tau pathologic features correlate with memory performance in patients with mild cognitive impairment. [Research Support, Non-U.S. Gov't]. *AJNR. American Journal of Neuroradiology, 29*(6), 1183–1189. doi:10.3174/ajnr.A1059.

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.

Jack, C. R., Jr., Bernstein, M. A., Borowski, B. J., Gunter, J. L., Fox, N. C., Thompson, P. M., et al. (2010a). Update on the magnetic resonance imaging core of the Alzheimer's disease neuroimaging initiative. *Alzheimers Dement, 6*(3), 212–220. doi:10.1016/j.jalz.2010.03.004.

Jack, C. R., Jr., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., et al. (2010b). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurology, 9*(1), 119–128.

Jack, C. R., Jr., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., et al. (2008). The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J Magn Reson Imaging, *27*(4), 685-691. doi:10.1002/jmri.21049.

Llano, D. A., Laforet, G., & Devanarayan, V. (2011). Derivation of a new ADAS-cog composite using tree-based multivariate analysis: prediction of conversion from mild cognitive impairment to Alzheimer disease. *Alzheimer Disease and Associated Disorders, 25*(1), 73–84.

McDonald, R. P. (1999). *Test theory: a unified treatment*. Mahwah: Erlbaum.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology, 34*(7), 939–944.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*: Routledge.

Mohs, R. C., Knopman, D., Petersen, R. C., Ferris, S. H., Ernesto, C., Grundman, M., et al. (1997). Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the Alzheimer's Disease Assessment Scale that broaden its scope. The Alzheimer's Disease Cooperative Study. *Alzheimer Disease and Associated Disorders, 11*(Suppl 2), S13–21.

Morris, J. C. (1993). The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology, 43*(11), 2412–2414.

Mungas, D., & Reed, B. R. (2000). Application of item response theory for development of a global functioning measure of dementia with linear measurement properties. *Statistics in Medicine, 19*(11–12), 1631–1644.

Murphy, E. A., Holland, D., Donohue, M., McEvoy, L. K., Hagler, D. J., Jr., Dale, A. M., et al. (2010). Six-month atrophy in MTL structures is associated with subsequent memory decline in elderly controls. [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't]. *NeuroImage, 53*(4), 1310–1317. doi:10.1016/j.neuroimage.2010.07.016.

Muthén, L., & Muthén, B. (2006). *Mplus users guide. Version 4.1 ed*. Los Angeles: Muthen and Muthen.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., et al. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care, 45*(5 Suppl 1), S22–31.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of life research: an international journal of quality of life aspects of treatment, care and rehabilitation, 16 Suppl 1*, 19–31.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552–66.

Rey, A. (1964). *L'examen clinique en psychologie*. Paris: Presses Universitaires de France.

Van Petten, C., Plante, E., Davidson, P. S., Kuo, T. Y., Bajuscak, L., & Glisky, E. L. (2004). Memory and executive function in older adults: relationships with temporal and prefrontal gray matter volumes and white matter hyperintensities. [Clinical Trial Research

Support, U.S. Gov't, P.H.S.]. *Neuropsychologia, 42*(10), 1313–1335. doi:10.1016/j.neuropsychologia. 2004.02.009.

Walhovd, K. B., Fjell, A. M., Amlien, I., Grambaite, R., Stenset, V., Bjornerud, A., et al. (2009). Multimodal imaging in mild cognitive impairment: metabolism, morphometry and diffusion of the temporal-parietal memory network. *NeuroImage, 45*(1), 215–223. doi:10.1016/j.neuroimage.2008.10.053.

Wechsler, D. (1987). *WMS-R: Wechsler Memory Scale—Revised manual*. NY: Psychological Corporation / HBJ.

Wouters, H., van Gool, W. A., Schmand, B., & Lindeboom, R. (2008). Revising the ADAS-cog for a more accurate assessment of cognitive impairment. *Alzheimer Disease and Associated Disorders, 22*(3), 236–244.

Yonelinas, A. P., Widaman, K., Mungas, D., Reed, B., Weiner, M. W., & Chui, H. C. (2007). Memory in the aging brain: doubly dissociating the contribution of the hippocampus and entorhinal cortex. *Hippocampus, 17*(11), 1134–1140. doi:10.1002/hipo.20341.