Published in final edited form as:

Psychol Assess. 2016 December ; 28(12): 1674–1683. doi:10.1037/pas0000293.

Factor Structure of the ARIC-NCS Neuropsychological Battery: An Evaluation of Invariance Across Vascular Factors and Demographic Characteristics

Andreea M. Rawlings¹, Karen Bandeen-Roche², Alden L. Gross¹, Rebecca F. Gottesman^{1,3},
Laura H. Coker⁴, Alan D. Penman⁵, A. Richey Sharrett¹, and Thomas H. Mosley⁶

¹Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland ²Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD ³Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, Maryland ⁴Department of Social Sciences and Health Policy, Wake Forest University School of Medicine, Medical Center Boulevard, Winston-Salem, NC ⁵Center of Biostatistics and Bioinformatics, University of Mississippi Medical Center, Jackson, MS ⁶Department of Medicine; University of Mississippi Medical Center, Jackson, MS

Abstract

Neuropsychological test batteries are designed to assess cognition in detail by measuring cognitive performance in multiple domains. This study examines the factor structure of tests from the ARIC-NCS battery overall and across informative subgroups defined by demographic and vascular risk factors in a population of older adults. We analyzed neuropsychological test scores from 6413 participants in the Atherosclerosis Risk in Communities Neurocognitive Study (ARIC-NCS) examined in 2011–2013. Confirmatory Factor Analysis (CFA) was used to assess the fit of an *a priori* hypothesized three-domain model, and fit statistics were calculated and compared to one- and two-domain models. Additionally, we tested for stability (invariance) of factor structures among different subgroups defined by diabetes, hypertension, age, sex, race, and education. Mean age of participants was 76 years, 76% were White, and 60% were female. CFA on the *a priori* hypothesized three-domain structure, including memory, sustained attention and processing speed, and language, fit the data better (CFI=0.973, RMSEA=0.059) than the two-domain (CFI=0.960, RMSEA=0.070) and one-domain (CFI=0.947, RMSEA=0.080) models. BIC value was lowest, and QQ-Plots indicated better fit, for the three-domain model. Additionally, multiple-group CFA supported a common structure across the tested demographic subgroups, and indicated strict invariance by diabetes and hypertension status. In this community-based population of older adults with varying levels of cognitive performance, the *a priori* hypothesized three-domain structure fit the data well. The identified factors were configurally invariant by age, sex, race, and education, and strictly invariant by diabetes and hypertension status.

Corresponding author: Andreea M. Rawlings, Welch Center for Prevention, Epidemiology and Clinical Research and the Johns Hopkins Bloomberg School of Public Health, 2024 E. Monument Street, Suite 2-600, Baltimore MD 21287, 443-287-4169 (phone) / 410-955-0476 (fax), arawlin2@jhu.edu.

Disclosures: none exist.

Keywords

adult; aged; cohort studies; factor analysis; statistical; cognition; neuropsychological tests

Introduction

Neuropsychological test batteries are designed to evaluate different areas of cognition in order to identify, track, and diagnose cognitive impairment and dementia (Hayden, Jones, et al., 2011). This motivates the grouping of neuropsychological tests to identify cognitive domains that may be differentially affected by disease pathology. Further, grouping cognitive tests into domains may reduce measurement error and may better facilitate testing of primary hypotheses regarding the possible etiology of various cognitive impairments (Gibbons, Bubb, & Brown, 2007; Silverstein, 2008).

Performance on neuropsychological tests may vary by age, global mental status, or for reasons unrelated to cerebral pathology, such as education level or cultural factors. It is important to determine if the underlying factors measured by cognitive tests are similar (invariant) across subgroups to ensure that observed variability in cognitive performance in different groups can be appropriately attributed to underlying cognitive abilities rather than to differences in the meaning of the tests. Such invariance analyses by demographics and cognitive status are important to establish and are commonly conducted (Hayden, Reed, et al., 2011; Mungas, Widaman, Reed, & Tomaszewski Farias, 2011; Park et al., 2012; Siedlecki, Honig, & Stern, 2008). Additionally, studies have examined invariance by genetic risk factors for Alzheimer's disease (AD) (Dowling, Hermann, La Rue, & Sager, 2010) with invariance among tests often varying by genetic risk factors in populations and by subgroups.

However, the examination of factor invariance by vascular risk factors, such as diabetes and hypertension, is less common. This is particularly important as hypertension and diabetes are common in the general population, and even more so in older adults and African Americans. Establishing the invariance of cognitive tests by vascular risk factors is vital for studies that aim to evaluate the associations of vascular risk factors or markers with cognitive decline and dementia.

Vascular dementia is the second most common type of dementia after AD, with the prevalence of pure vascular dementia about 10-15%. However, emerging evidence shows that most dementia may be a mix of both vascular dementia and AD pathology (O'Brien & Thomas, 2015). Studies have also shown that the clinical expression of dementia is larger in the presence of vascular disease, independent of the level of AD pathology (Jellinger & Attems, 2015; Snowden et al., 1997). Hypertension and diabetes have been identified as risk factors for vascular dementia (Gorelick, 2004) and may be contributors to AD (de Bruijn & Ikram, 2014). Both of these risk factors are very common in older adults and, more importantly, are modifiable, providing potential means of intervention to prevent or delay the vascular contributions to dementia and AD.

The present study addressed two aims. The first aim was to explore the factor structure of the Atherosclerosis Risk in Communities Neurocognitive Study (ARIC-NCS) neuropsychological test battery using confirmatory factor analysis (CFA) for *a priori* hypothesized domains. The second aim was to explore the stability of the factor structure by demographic subgroups (defined by age, race, sex, and education) and vascular risk factors (diabetes and hypertension). The ARIC-NCS battery contains commonly used tests, and is very similar to the neuropsychological test battery from the National Institute on Aging Uniform Data Set (Morris et al., 2006; Weintraub et al., 2009).

Method

Study Population

The Atherosclerosis Risk in Communities (ARIC) Study is a bi-ethnic, community-based prospective cohort of 15,792 middle aged adults from four U.S. communities: Washington County, Maryland; Forsyth County, North Carolina; suburbs of Minneapolis, Minnesota; and Jackson, Mississippi. ARIC participants were seen at four in-person visits roughly 3 years apart, from 1987-89 for visit 1, through 1996-98 for visit 4. A fifth visit took place from 2011-2013, which included neuropsychological testing as part of the ARIC-NCS. 6538 participants attended visit 5, and 6501 completed the neurocognitive assessment. We excluded participants who were neither Black nor White ($n=20$) or who were missing all cognitive tests ($n=68$), giving a sample size of 6413 for the present study. Institutional review boards at each study site reviewed and approved the study; written informed consent was obtained from all participants.

Neuropsychological Assessment

The ARIC-NCS test battery included eleven of the neuropsychological tests recommended for inclusion in the National Institute of Aging's National Alzheimer's Disease Coordinating Center (NACC) Uniform Data Set Battery (Morris et al., 2006). Protocols for the tests were standardized and examiners were trained centrally. The tests were administered in a fixed order during one session in a quiet room.

We hypothesized that the tests represented three cognitive domains: Memory, Language and Verbal Fluency, and Sustained Attention and Processing Speed (SAPS), and that this grouping of tests best represents the data. This *a priori* hypothesized 3-domain structure was based in part on the presumed underlying neurological structures involved (Lezak, 2012) as well as findings from two studies, which also used tests comprising the NACC Uniform Data Set Battery, namely the Alzheimer's Disease Neuroimaging Initiative (Park et al., 2012) and the NACC study (Hayden, Jones, et al., 2011). Because validation necessitates not only verifying that a three-domain model well-characterizes the observed data, but also that models of lower dimensions do not suffice to this end, we additionally examined a one-domain model (all tests grouped together) and a two-domain model (a memory domain and a domain of the remaining tests, representing a language/SAPS combined domain). The tests, grouped according to our *a priori* hypothesized three-domain structure, are described below.

Memory Domain

Delayed Word Recall Test (DWRT): In the DWRT participants are presented with 10 common nouns that they are asked to use in a sentence. Two exposures to the words are given. After a five-minute delay, participants are given 60 seconds to recall the words. The score for the DWRT is the number of words correctly recalled.

Logical Memory Test (LMT): In part 1, participants are read two short stories and are asked to recall the details immediately following each story. At the conclusion of Part 1, participants are told they will be asked again about the stories. Part 2 is completed after a filled delay of approximately 20 minutes, and consists of participants recalling details of the same stories from part 1. The metric for both parts is the number of correct details recalled, with a maximum score of 50 for each of the two parts. Here we present parts 1 and 2 separately, labelled as “LM 1” and “LM 2”, respectively.

Incidental Learning: Incidental learning is based on the Digit Symbol Substitution Test (DSST). For the DSST, participants are not instructed to learn the digit-symbol pairs. Immediately following completion of the DSST, participants are asked to remember the symbols and corresponding digit-symbol pairs. The metric used for this test is the number of symbol-pairs correctly recalled, with a range of 0 to 9.

Language and Verbal Fluency Domain

Animal Naming: In this test, participants are asked to name as many animals as they can in 60 seconds. Names of extinct, imaginary, and magical animals are admissible. Credit also is given for breeds, different names for males, females, or infants of the same species (e.g. bull, cow, calf) as well as superordinate and subordinate (e.g. dog and terrier) for birds, reptiles, and insects. The score is the total number of animals generated.

Boston Naming Test (BNT): In the BNT, participants are shown a series of 30 line drawings, one at a time, and are given 20 seconds to name the object shown in each drawing. No hints or clues are provided. The variable used in our analysis is the number of drawings correctly identified.

Word Fluency Test (WFT): The WFT is a test of executive function and language. Participants are given 60 seconds to generate as many words as possible beginning with the letters F, A and S (60 seconds for each letter), avoiding proper nouns. The WFT score is the total number of acceptable words generated for the three letters

Sustained Attention and Processing Speed (SAPS) Domain

Trail Making Test (TMT): The TMT is comprised of two parts, A and B. In part A, participants are presented with numbers 1-25 each in a separate circle and distributed haphazardly across a page, and are asked to draw lines connecting the numbers sequentially. Similarly in part B, participants are given a page containing numbers (1-13) and letters (A-L), and are asked to draw lines connecting the numbers and letters in sequential, but alternating, fashion. Time to completion is the metric used for both parts. Participant who

take longer than four minutes to complete the test, or who make more than 5 errors, are given a maximum time of 240 seconds.

Digit Symbol Substitution Test (DSST): For the DSST, from the Wechsler Adult Intelligence Scale-revised (WAIS-R), participants are asked to translate numbers to symbols using a key. The score is the total number of numbers correctly translated to symbols within 90-seconds and the range of possible scores is 0 to 93.

Digit Span Backwards (DSB): In the DSB, participants are read a series of numbers increasing in length from 2 to 7 digits each and are asked to repeat each series backwards. There are two trials for each digit span length, giving a maximum score of 12.

Hypertension Assessment—Blood pressure was measured using an OMRON HEM-907XL automated blood pressure monitor. Following a five-minute quiet rest period, three blood pressure measurements were taken, and the second and third measurements were averaged. Hypertension was defined as an average (of the second and third readings) systolic blood pressure greater than 140, a diastolic blood pressure greater than 90, or self-reported blood-pressure lowering medication use.

Diabetes Assessment

Participants who had a measured haemoglobin A1c $\geq 6.5\%$, or who brought glucose-lowering medication to the study visit were classified as having diabetes. Additionally, participants who self-reported a diagnosis of diabetes or glucose-lowering medication use during annual follow-up telephone calls prior to the study visit were also classified as having diabetes.

Statistical Analysis

For each cognitive test, we calculated standardized Z scores by subtracting the test mean from each participant's test score and dividing by the test standard deviation. For the Trail Making tests, Z scores were calculated after taking the log of the test scores. In addition, because a higher score on these tests indicates worse performance, we multiplied the Z score by -1 so that low Z scores indicated worse performance for all tests. Participants who did not complete a test due to difficulty were assigned a Z score of -2 as described in ARIC-NCS Manual 17 ("Manual 17. ARIC Neurocognitive Exam (Stages 2 and 3)," 2011).

Cronbach's alpha (Cronbach, 1951) was used to describe internal consistency reliability for test scores of each domain of the *a priori* hypothesized three-domain structure.

Factor Analysis

We examined the domain (factor) structure of the cognitive tests using the test-specific Z scores and confirmatory factor analysis (CFA). CFA was applied to the *a priori* hypothesized three-domain model (tests grouped as described above), the two-domain model, and the one-domain model. CFA models were fit using full-information maximum likelihood, which allows the inclusion of all participants who have at least one test score. Correlations between the errors of the two logical memory tests and the two trail making tests were included based

on *a priori* expectation that the errors of these pairs of tests would be correlated as the tests themselves are highly related. The correlation between the trail making test part A and the digit symbol substitution test was included based on examination of model fit. Including these correlations allowed us to relax the assumption of conditional independence and resulted in improved model fit. Analyses were completed using Stata/SE 13.1 (StataCorp LP, College Station TX).

Configural invariance is met when the same tests are associated with the same factors in each group, and is evaluated using model fit statistics. We focused on configural invariance for subgroups defined by age (dichotomized at the median age), sex, race, and education, using the three-domain structure. Education was grouped into three levels: less than high school (<HS), high school or vocational school (HS), or more than high school (>HS, includes any college or professional school). For hypertension and diabetes, we further examined metric, strong, and strict invariance. Metric invariance is met when factor loadings do not differ between subgroups. Strong invariance is met when factor loadings and intercepts do not differ between subgroups. Finally, strict invariance is met when factor loadings, intercepts, and residual variances do not differ between subgroups. The three error correlations (between the logical memory tests, trail making tests, trails part A and digit symbol substitution) were unconstrained (allowed to vary) between subgroups for configural invariance, but were constrained for metric, strong, and strict invariance.

We examined five model fit statistics to assess both relative and absolute model fit. The Comparative Fit Index (CFI)(Bentler & Mooijaart, 1989; Bentler, 1990) compares the fitted model with a null model that assumes uncorrelated variables (i.e. the independence model). The Tucker-Lewis Index (TLI)(Marsh & Hau, 1996; Tucker & Lewis, 1973) is a similar measure of relative model fit that indicates improvement over the null model. CFI and TLI values range from 0-1, with values >0.9 indicating good model fit. The Root Mean Square Error of Approximation (RMSEA)(M. W. Browne & Cudeck, 1992; Steiger, 1989, 1990) and the Standardized Root Mean Residual (SRMR)(Steiger, 1989) assess absolute model fit. They are measures of the size of the model residuals and are insensitive to sample size and variable distribution. For RMSEA, values <0.05 indicate very good fit. For SRMR, values <0.08 indicate adequate fit, and values <0.05 indicate good fit. The Bayesian Information Criterion (BIC) (Raftery, 1995), is a criterion for model selection that adds a penalty for increasing model complexity and possible over-fitting of the data; the preferred model is the one with the lowest BIC. Additionally, changes in BIC values can be interpreted using grades of evidence described by (Raftery, 1995), where decreases in BIC >10 indicate very strong evidence to prefer the model with the lower BIC. Finally, for the one-, two-, and three-domain models we calculated the correlation matrix residuals and created quantile-quantile (QQ) plots for the residuals of one-versus two-domain models and the two- versus three-domain models to further evaluate relative model fit (Michael W Browne, MacCallum, Kim, Andersen, & Glaser, 2002).

Results

Characteristics of study participants and the distributions of the raw, non-standardized test scores are shown in Table 1. Participants' mean age was approximately 76 years, 76% were

White, and 59% were female. The prevalence of vascular risk factors was high, with 74% of participants having hypertension, 33% diabetes, and 15% reporting a history of coronary heart disease.

The *a priori* hypothesized three-domain model with standardized factor loadings, correlations between the factors, and residual errors are shown in Figure 1. Correlations between the factors were relatively high, with values of 0.80, 0.82, and 0.85 for Memory/SAPS, Memory/Language, and Language/SAPS, respectively.

Fit statistics for the one-, two-, and three-domain models are shown in Table 2. Fit statistics for the one-domain model indicated good absolute fit (CFI=0.947, RMSEA=0.080), however they indicated worse fit than the two-domain model (CFI=0.961, RMSEA=0.070); the BIC value for the one-domain model was highest among the models considered. All fit statistics indicated better fit for the three-domain model (CFI=0.973, RMSEA=0.059), compared to the one- and two-domain models, and confidence interval for RMSEA excluded the confidence intervals for RMSEA from one and two-domain models. Of note, DSB had the lowest factor loading compared to all other tests, and compared to tests within SAPS. Thus, in a CFA where we excluded this test from our analyses, model fit statistics indicated even better fit (CFI=0.981, RMSEA=0.056). Figure 2 shows the QQ plots for the correlation matrix residuals from models for two-domains versus one-domain (Figure 2, Panel A) and three-domains versus two-domains (Figure 2, Panel B). Quantiles of the residuals indicate superior fit for the three-domain model compared to the two- and one-domain models.

Tables 3 and 4 show standardized factor loadings and fit statistics for the subgroup analysis of diabetes and hypertension, respectively, by each model of invariance. For both subgroups, fit statistics across all four models were similar and did not suggest deterioration in model fit as parameters were restricted to be similar across subgroups. Additionally, BIC values were lowest for the model with strict invariance; this model additionally had a BIC value that was more than 10 units lower than the other models, giving strong evidence to prefer it to the other models.

Standardized factor loadings and fit statistics of our demographic subgroup analyses are shown in Table 5. Model fit statistics for multiple group models based on age, sex, education and race were similar to the overall model, indicating configural invariance for these subgroups. RMSEA values ranged from 0.057 to 0.061 and SRMR values ranged from 0.033 to 0.037, both indicating configural invariance in each subgroup. Further exploration of invariance across these demographic factors indicated at least metric invariance across all demographic factors (Online Tables 1-4).

The internal consistency reliability of each domain was good, with alpha values of 0.81, 0.72, and 0.78 for the memory, language, and SAPS, respectively. Test scores within each domain had similar correlations with the domain, indicating consistency. For example, Animal Naming, Boston Naming, and Word Fluency test scores were correlated 0.827, 0.785 and 0.806 respectively with the Language and Verbal Fluency domain. An exception was scores of the DSB, which were less correlated with the SAPS domain than were test scores of Trails A, Trails B, and DSS (Online Table 5).

Discussion

The growing interest in clarifying structural and functional associations in aging and disease in the context of the rapidly expanding minority and aging populations in the US motivates the need to demonstrate that the measures commonly used in clinical and epidemiologic studies have a stable structure (i.e., reflect the same construct) across potentially informative subgroups of vascular risk factors and demographics. Our *a priori* hypothesized three-domain structure based on our expectations from available published evidence (i.e. (1) Memory, (2) Language and Verbal Fluency, and (3) Sustained attention and Processing Speed), fit the data better than a one- or two-domain model. Additionally, we have established invariance in a diverse, community-based population of older adults, using a cognitive battery comprised of common tests. Our analyses of the stability of the factors across subgroups indicated configural invariance, meaning the same domains are being measured regardless of age, race, sex, or education. Further, examination of invariance by diabetes and hypertension status suggested substantial stability by these vascular risk factors. Similar domain structures and invariance between demographic factors have been previously reported (Dowling et al., 2010; Hayden, Jones, et al., 2011; Jack Jr. et al., 2012; Mungas et al., 2011; Siedlecki et al., 2008; Vemuri et al., 2012), however few studies have examined invariance by vascular risk factors.

Establishing invariance by vascular risk factors is particularly important, as diabetes and hypertension are common, especially so among Blacks and older adults. The ability to identify meaningful dimensions of cognitive function is relevant not only for diagnostic purposes but also for characterizing those at highest risk for cognitive decline and dementia (e.g., by race or other demographics) and informing underlying brain functional-structural relationships. Establishing invariance for these very common vascular risk factors in a diverse sample is an important prerequisite for addressing these types of questions.

Our study has both limitations and strengths. The first limitation is that we have only 3-4 tests to represent each cognitive domain, thus we may have limited precision to fully characterize the domains. For example, DSB had the lowest factor loading compared to the other tests in the SAPS domain. This may indicate that DSB more finely measures an attention construct, whereas the other tests included in this domain relate more to processing speed and executive function. Second, Blacks in ARIC were recruited primarily from two communities, which may limit generalizing to other regions. However, the education backgrounds in Blacks from these centers are diverse, and the finding that the domain structure appears similar is encouraging. We note that while the alphas we estimated in our study are large enough for the purpose of making group-level comparisons, they are not large enough to make individual-level inferences (Nunnally & Bernstein, 1994).

A key strength of our study is the administration of a core of widely used tests, using a standardized protocol of centrally trained testers, to this large, diverse, community-based population. The battery of tests administered in ARIC-NCS is consistent with other large-scale studies, which provides for comparability across studies (Hayden, Jones, et al., 2011; Park et al., 2012; Siedlecki et al., 2008). Additionally, the use of full information maximum likelihood in the CFA analyses allowed us to include all participants. Lastly, the ability to

infer appropriate conclusions about group differences on neuropsychological test performance presumes careful attention to the selection of tests, availability of relevant norms for comparison, and sensitivity to the measurement process/testing situation that may differentially impact performance. Our study goes one step further in addressing the robustness of the construct validity of the underlying cognitive domains under study.

The choice of the number of factors in a given factor analysis depends on a number of substantive considerations in addition to statistical fit. Previous studies have noted that persons with normal cognition tend to have less variability in their performance and a one-domain model is typically found (Gross, Jones, Fong, Tommet, & Inouye, 2014; Jones et al., 2010; Strauss & Fritsch, 2004). In contrast, individuals with impairment tend to have more heterogeneity in their performance and more than one factor may be needed to accurately capture status distinctions because different cognitive abilities may deteriorate at different rates (Bakkour, Morris, Wolk, & Dickerson, 2013; Hayden, Reed, et al., 2011; Kanne, Balota, Storandt, McKeel, & Morris, 1998). In our study there was a persuasive improvement in fit for a three-factor model as compared to models with fewer factors or domains. Further, BIC values, which penalize for over-fitting, also indicated a three-domain model. However, we observed high correlations between the domains, ranging from 0.80-0.85, and the fit of the one-domain model was reasonable. Thus, using a global composite may have the advantage of greater overall reliability for characterizing individuals compared with composites of fewer tests, and perhaps greater sensitivity to the effects of the key causal factors that will be assessed in ARIC-NCS. The decision of the number of factors thus depends to some extent on the ultimate goal of the factor analysis.

In summary, in this community-based population of older adults, results from the CFA indicate that the ARIC-NCS battery may be summarized into three domains, and that these domains are stable across subgroups defined by age, race, education, diabetes or hypertension. For investigations into the contributions of vascular predictors (such as hypertension and diabetes) to cognition in older persons, it is vital to establish the invariance of cognitive domains by these risk factors. Our findings assure us, and future investigators, that one can use the same cognitive domain constructs to pursue questions regarding the vascular contribution to impairments. Using multiple indicators of cognitive performance may allow us to capture the general determinants of cognitive function, and average out method-specific and error components.

These results provide compelling evidence for the robustness of cognitive domains measured by our test battery. Our findings are encouraging for studies aiming to test hypotheses regarding the associations between midlife vascular factors and late-life cognitive impairment in diverse populations defined by age, race, education, and vascular risk factors.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). Neurocognitive data is collected by U01 HL096812, HL096814, HL096899, HL096902, HL096917 with previous brain MRI examinations funded by R01-HL70825. The authors thank the staff and participants of the ARIC study for their important contributions.

References

- Bakkour A, Morris JC, Wolk DA, Dickerson BC. The effects of aging and Alzheimer's disease on cerebral cortical anatomy: specificity and differential relationships with cognition. *NeuroImage*. 2013; 76:332–44. DOI: 10.1016/j.neuroimage.2013.02.059 [PubMed: 23507382]
- Bentler PM. Comparative fit indexes in structural models. *Psychological Bulletin*. 1990; 107(2):238–46. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/2320703>. [PubMed: 2320703]
- Bentler PM, Mooijjaart A. Choice of structural model via parsimony: a rationale based on precision. *Psychological Bulletin*. 1989; 106(2):315–7. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/2678203>. [PubMed: 2678203]
- Browne MW, Cudeck R. Alternative Ways of Assessing Model Fit. *Sociological Methods & Research*. 1992; 21(2):230–258. DOI: 10.1177/0049124192021002005
- Browne MW, MacCallum RC, Kim CT, Andersen BL, Glaser R. When fit indices and residuals are incompatible. *Psychological Methods*. 2002; 7(4):403–21. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2435310&tool=pmcentrez&rendertype=abstract>. [PubMed: 12530701]
- Cronbach L. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951; 16(3):297–334. DOI: 10.1007/bf02310555
- de Bruijn RFAG, Ikram MA. Cardiovascular risk factors and future risk of Alzheimer's disease. *BMC Medicine*. 2014; 12:130.doi: 10.1186/s12916-014-0130-5 [PubMed: 25385322]
- Dowling NM, Hermann B, La Rue A, Sager MA. Latent structure and factorial invariance of a neuropsychological test battery for the study of preclinical Alzheimer's disease. *Neuropsychology*. 2010; 24(6):742–756. DOI: 10.1037/a0020176 [PubMed: 21038965]
- Gibbons S, Bubb R, Brown B. Reducing Error: Averaging Data to Determine Factor Structure of the QMPR. 2007:3348–3355.
- Gorelick PB. Risk factors for vascular dementia and Alzheimer disease. *Stroke; a Journal of Cerebral Circulation*. 2004; 35(11 Suppl 1):2620–2. DOI: 10.1161/01.STR.0000143318.70292.47 [PubMed: 15375299]
- Gross AL, Jones RN, Fong TG, Tommet D, Inouye SK. Calibration and validation of an innovative approach for estimating general cognitive performance. *Neuroepidemiology*. 2014; 42(3):144–53. DOI: 10.1159/000357647 [PubMed: 24481241]
- Hayden KM, Jones RN, Zimmer C, Plassman BL, Browndyke JN, Pieper C, et al. Welsh-Bohmer KA. Factor structure of the National Alzheimer's Coordinating Centers uniform dataset neuropsychological battery: an evaluation of invariance between and within groups over time. *Alzheimer Dis Assoc Disord*. 2011; 25(2):128–137. DOI: 10.1097/WAD.0b013e3181ffa76d [PubMed: 21606904]
- Hayden KM, Reed BR, Manly JJ, Tommet D, Pietrzak RH, Chelune GJ, et al. Jones RN. Cognitive decline in the elderly: an analysis of population heterogeneity. *Age Ageing*. 2011; 40(6):684–689. DOI: 10.1093/ageing/afr101 [PubMed: 21890481]
- Jack CR Jr, Knopman DS, Weigand SD, Wiste HJ, Vemuri P, Lowe V, et al. Petersen RC. An operational approach to National Institute on Aging-Alzheimer's Association criteria for preclinical Alzheimer disease. *Ann Neurol*. 2012; 71(6):765–775. DOI: 10.1002/ana.22628 [PubMed: 22488240]
- Jellinger KA, Attems J. Challenges of multimorbidity of the aging brain: a critical update. *Journal of Neural Transmission (Vienna, Austria : 1996)*. 2015; 122(4):505–21. DOI: 10.1007/s00702-014-1288-x

- Jones RN, Rudolph JL, Inouye SK, Yang FM, Fong TG, Milberg WP, et al. Marcantonio ER. Development of a unidimensional composite measure of neuropsychological functioning in older cardiac surgery patients with good measurement precision. *Journal of Clinical and Experimental Neuropsychology*. 2010; 32(10):1041–9. DOI: 10.1080/13803391003662728 [PubMed: 20446144]
- Kanne SM, Balota DA, Storandt M, McKeel DW, Morris JC. Relating anatomy to function in Alzheimer's disease: neuropsychological profiles predict regional neuropathology 5 years later. *Neurology*. 1998; 50(4):979–85. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/9566382>. [PubMed: 9566382]
- Lezak, MD. *Neuropsychological assessment*. 5th. Oxford; New York: Oxford University Press; 2012.
- Manual 17. ARIC Neurocognitive Exam (Stages 2 and 3). 2011. Retrieved from <http://www2.csc.unc.edu/aric/>
- Marsh HW, Hau KT. Assessing Goodness of Fit: Is Parsimony Always Desirable? *The Journal of Experimental Education*. 1996; 64(4):364–390. DOI: 10.1080/00220973.1996.10806604
- Morris JC, Weintraub S, Chui HC, Cummings J, Decarli C, Ferris S, et al. Kukull WA. The Uniform Data Set (UDS): clinical and cognitive variables and descriptive data from Alzheimer Disease Centers. *Alzheimer Dis Assoc Disord*. 2006; 20(4):210–216. DOI: 10.1097/01.wad.0000213865.09806.92 [PubMed: 17132964]
- Mungas D, Widaman KF, Reed BR, Tomaszewski Farias S. Measurement invariance of neuropsychological tests in diverse older persons. *Neuropsychology*. 2011; 25(2):260–9. DOI: 10.1037/a0021090 [PubMed: 21381830]
- Nunnally, JC.; Bernstein, IH. *Psychometric Theory*. 3rd. New York, NY: McGraw-Hill; 1994.
- O'Brien JT, Thomas A. Vascular dementia. *The Lancet*. 2015; 386(10004):1698–1706. DOI: 10.1016/S0140-6736(15)00463-8
- Park LQ, Gross AL, McLaren DG, Pa J, Johnson JK, Mitchell M, et al. Alzheimer's Disease Neuroimaging, I. Confirmatory factor analysis of the ADNI Neuropsychological Battery. *Brain Imaging Behav*. 2012; 6(4):528–539. DOI: 10.1007/s11682-012-9190-3 [PubMed: 22777078]
- Raftery AE. Bayesian Model Selection in Social Research. *Sociological Methodology*. 1995; doi: 10.2307/271063
- Siedlecki KL, Honig LS, Stern Y. Exploring the structure of a neuropsychological battery across healthy elders and those with questionable dementia and Alzheimer's disease. *Neuropsychology*. 2008; 22(3):400–411. DOI: 10.1037/0894-4105.22.3.400 [PubMed: 18444718]
- Silverstein SM. Measuring specific, rather than generalized, cognitive deficits and maximizing between-group effect size in studies of cognition and cognitive change. *Schizophrenia Bulletin*. 2008; 34(4):645–55. DOI: 10.1093/schbul/sbn032 [PubMed: 18468987]
- Snowdon DA, Greiner LH, Mortimer JA, Riley KP, Greiner PA, Markesbery WR. Brain infarction and the clinical expression of Alzheimer disease. The Nun Study. *JAMA*. 1997; 277(10):813–7. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/9052711>. [PubMed: 9052711]
- Steiger, JH. EZPATH: A Supplementary module for SYSTAT and SYGRAPH. Evanston, IL: SYSTAT; 1989.
- Steiger JH. Structural Model Evaluation and Modification: An Interval Estimation Approach. *Multivariate Behavioral Research*. 1990; 25(2):173–180. DOI: 10.1207/s15327906mbr2502_4 [PubMed: 26794479]
- Strauss ME, Fritsch T. Factor structure of the CERAD neuropsychological battery. *Journal of the International Neuropsychological Society: JINS*. 2004; 10(4):559–65. DOI: 10.1017/S1355617704104098 [PubMed: 15327734]
- Tucker LR, Lewis C. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*. 1973; 38(1):1–10. DOI: 10.1007/BF02291170
- Vemuri P, Lesnick TG, Przybelski SA, Knopman DS, Roberts RO, Lowe VJ, et al. Jack CR Jr. Effect of lifestyle activities on Alzheimer disease biomarkers and cognition. *Ann Neurol*. 2012; 72(5): 730–738. DOI: 10.1002/ana.23665 [PubMed: 23280791]
- Weintraub S, Salmon D, Mercaldo N, Ferris S, Graff-Radford NR, Chui H, et al. Morris JC. The Alzheimer's Disease Centers' Uniform Data Set (UDS): the neuropsychologic test battery.

Alzheimer Disease and Associated Disorders. 2009; 23(2):91–101. DOI: 10.1097/WAD.0b013e318191c7dd [PubMed: 19474567]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1. Confirmatory factory analysis model derived from *a priori* hypothesized domain structure

Latent factors are shown in ovals, measured tests are shown in rectangles, and error terms are shown in circles. Values shown between factors and between factors and tests are correlations. Values along curved and straight arrows are correlations, while values shown outside the circles for ϵ_1 - ϵ_{11} are residual variances.

Figure 2. Quantile-quantile plot of correlation matrix residuals comparing one- versus two-domain models and two- versus three-domain models

Each model fit implies estimates for all correlations among neuropsychological tests.

Residuals from model fits (observed-estimated) are shown as qq-plots in which percentiles of the residuals from the two models are plotted against one another (axes are labeled with original units of the residuals). Equivalent fits appear as equivalent distributions of the deviations between fitted and observed values, hence lie along the $y=x$ line (shown as solid line). Panel A: residuals from the one-domain model plotted versus residuals from the two-domain model. Panel B: residuals from the two-domain model plotted versus residuals from the three-domain model. In both panels, there is systematic deviation from the $y=x$ line with residuals from models with lower numbers of domains more dispersed than those from models with higher numbers of domains, indicating inferior fit.

Table 1
Characteristics of study participants, N=6413

	Total	25th, 75th percentile
Age, mean (SD)	76.2 (5.2)	71.9, 80.0
White, N (%)	4900 (76.4)	-
Female, N (%)	3770 (58.8)	-
Education, N (%)		
Less than high school	956 (14.9)	-
High school or GED	2672 (41.7)	-
College or vocational school	2774 (43.3)	-
Hypertension, N (%)	4744 (74.0)	-
Diabetes, N (%)	2091 (32.6)	-
History of coronary heart disease, N (%)	944 (14.7)	-
Cognitive Tests *		
Delayed Word Recall, words recalled	5.2 (1.9)	4, 6
Logical Memory 1, details recalled	21.5 (7.5)	16, 27
Logical Memory 2, details recalled	16.6 (7.9)	11, 22
Incidental Learning, symbol-pairs recalled	3.3 (2.3)	1, 5
Trails A, time (seconds)	50.2 (31.1)	33, 56
Trails B, time (seconds)	128.3 (60.6)	81, 165
Digit Symbol Substitution, symbols translated	37.8 (12.1)	29, 46
Digit Span Backwards, correct spans	5.5 (2.0)	4, 7
Animal Naming, animals generated	16.0 (5.1)	13, 19
Boston Naming, items identified	24.6 (5.5)	23, 28
Word Fluency, words generated	32.6 (12.4)	24, 41

* Sample size for individual tests vary; values reported as mean (SD)

Table 2
Fit statistics from confirmatory factor analyses by number of domains and domain definitions

	1 domain	2 domains	3 domains (all tests)	3 domains (without DSB)
CFI	0.947	0.961	0.973	0.981
TLI	0.929	0.946	0.962	0.971
RMSEA *	0.080 (0.077, 0.084)	0.070 (0.067, 0.073)	0.059 (0.056, 0.063)	0.056 (0.052, 0.060)
SRMR [†]	0.044	0.046	0.033	0.029
BIC	160,804	160,372	159,985	143,953

* RMSEA listed as estimate (90% confidence interval)

[†] SRMR calculated from models restricted to complete data

Abbreviations: DSB, digit span backwards; CFI, comparative fit index (>0.90 indicates good fit); TLI, Tucker-Lewis index(>0.90 indicates good fit); RMSEA, root mean squared error of approximation(<0.10 indicates good fit, <0.05 indicates very good fit); SRMR, standardized root mean squared residual (<0.08 indicates adequate fit, <0.05 indicates good fit); BIC, Bayesian information criterion (lower numbers are better, and decreases >10 indicate strong evidence to prefer the model with the lower BIC).

Table 3
Standardized factor loadings and fit statistics for invariance models by diabetes

	Configural Invariance Variances constrained to 1, means to 0; intercepts, residuals, and factor loadings unconstrained		Metric Invariance Variances constrained to 1, means to 0, factor loadings invariant across groups; intercepts and residuals unconstrained		Strong Invariance Variances constrained to 1, means to 0, factor loadings and intercepts invariant across groups; residuals unconstrained		Strict Invariance Variances constrained to 1, means to 0, factor loadings, intercepts, and residuals invariant across groups	
	No Diabetes	Diabetes	No Diabetes	Diabetes	No Diabetes	Diabetes	No Diabetes	Diabetes
Memory								
DWR	0.64	0.66	0.64	0.65	0.64	0.65	0.64	0.65
LM 1	0.71	0.70	0.70	0.71	0.70	0.70	0.70	0.71
LM 2	0.73	0.73	0.73	0.73	0.73	0.73	0.72	0.74
Incidental Learning	0.63	0.67	0.63	0.67	0.63	0.67	0.64	0.65
Language								
Animal Naming	0.73	0.72	0.72	0.74	0.72	0.74	0.72	0.73
Word Fluency	0.65	0.70	0.65	0.69	0.66	0.70	0.67	0.68
Boston Naming	0.65	0.68	0.66	0.66	0.66	0.66	0.66	0.67
SAPS								
Trails A	0.67	0.72	0.68	0.69	0.68	0.69	0.69	0.69
Trails B	0.83	0.83	0.84	0.83	0.83	0.83	0.83	0.83
DSS	0.79	0.81	0.79	0.81	0.79	0.81	0.80	0.80
DSB	0.51	0.53	0.51	0.53	0.50	0.53	0.51	0.51
Fit statistics								
CFI	0.973		0.973		0.971		0.970	
TLI	0.961		0.964		0.965		0.969	
RMSEA *	0.060 (0.056, 0.063)		0.057 (0.054, 0.060)		0.056 (0.053, 0.059)		0.053 (0.050, 0.056)	
SRMR †	0.033		0.034		0.035		0.036	
BIC	158,916		158,868		158,852		158,783	

* RMSEA listed as value (90% confidence interval)

† SRMR calculated from models restricted to complete data

Abbreviations: DWR, delayed word recall; LM, logical memory; SAPS, sustained attention and processing speed; DSS, digit span backwards; CFI, comparative fit index (>0.90 indicates good fit); TLI, Tucker-Lewis index (>0.90 indicates good fit); RMSEA, root mean squared error of approximation (<0.10 indicates good fit, <0.05 indicates very good fit); SRMR, standardized root mean squared residual (<0.08 indicates adequate fit, <0.05 indicates good fit); BIC, Bayesian information criterion (lower numbers are better, and decreases >10 indicate strong evidence to prefer the model with the lower BIC).

Table 4
Standardized factor loadings and fit statistics for invariance models by hypertension

	Configural Invariance Variances constrained to 1, means to 0; intercepts, residuals, and factor loadings unconstrained		Metric Invariance Variances constrained to 1, means to 0, factor loadings invariant across groups; intercepts and residuals unconstrained		Strong Invariance Variances constrained to 1, means to 0, factor loadings and intercepts invariant across groups; residuals unconstrained		Strict Invariance Variances constrained to 1, means to 0, factor loadings, intercepts, and residuals invariant across groups	
	No Htn	Htn	No Htn	Htn	No Htn	Htn	No Htn	Htn
<u>Memory</u>								
DWR	0.67	0.63	0.66	0.64	0.66	0.64	0.66	0.64
LM 1	0.71	0.70	0.71	0.70	0.71	0.70	0.71	0.69
LM 2	0.73	0.72	0.74	0.72	0.74	0.72	0.74	0.72
Incidental Learning	0.63	0.65	0.64	0.64	0.64	0.64	0.66	0.64
<u>Language</u>								
Animal Naming	0.74	0.72	0.73	0.73	0.73	0.73	0.74	0.72
Word Fluency	0.65	0.68	0.66	0.67	0.66	0.67	0.68	0.67
Boston Naming	0.65	0.67	0.65	0.67	0.65	0.67	0.67	0.66
<u>SAPS</u>								
Trails A	0.69	0.69	0.70	0.68	0.70	0.68	0.69	0.68
Trails B	0.85	0.83	0.86	0.82	0.86	0.83	0.84	0.83
DSS	0.78	0.80	0.79	0.80	0.79	0.80	0.80	0.80
DSB	0.52	0.51	0.50	0.52	0.50	0.51	0.52	0.51
<u>Fit statistics</u>								
CFI	0.973		0.973		0.973		0.971	
TLI	0.961		0.965		0.967		0.970	
RMSEA *	0.059 (0.056, 0.063)		0.056 (0.053, 0.060)		0.055 (0.052, 0.058)		0.052 (0.049, 0.055)	
SRMR †	0.034		0.034		0.036		0.036	
BIC	158,441		158,378		158,342		158,277	

* RMSEA listed as value (90% confidence interval)

† SRMR calculated from models restricted to complete data

Abbreviations: Htn, hypertension; DWR, delayed word recall; LM, logical memory; SAPS, sustained attention and processing speed; DSS, digit symbol substitution; DSB, digit span backwards; CFI, comparative fit index (>0.90 indicates good fit); TLI, Tucker-Lewis index (>0.90 indicates good fit); RMSEA, root mean squared error of approximation (<0.10 indicates good fit, <0.05 indicates very good fit); SRMR, standardized root mean squared residual (<0.08 indicates adequate fit, <0.05 indicates good fit); BIC, Bayesian information criterion (lower numbers are better, and decreases >10 indicate strong evidence to prefer the model with the lower BIC).

Table 5
Standardized factor loadings and fit statistics by subgroups of age, sex, race, and education

Domain/Test	Age		Sex		Race		Education	
	<75	75	Male	Female	White	Black	<HS	>HS
<u>Memory</u>								
DWR	0.57	0.65	0.62	0.67	0.64	0.68	0.68	0.66
LM 1	0.62	0.73	0.68	0.72	0.69	0.74	0.67	0.65
LM 2	0.65	0.75	0.71	0.74	0.72	0.75	0.70	0.69
Incidental Learning	0.60	0.65	0.65	0.64	0.64	0.65	0.59	0.64
<u>Language</u>								
Animal Naming	0.67	0.73	0.71	0.74	0.73	0.74	0.70	0.76
Word Fluency	0.71	0.67	0.68	0.66	0.64	0.77	0.73	0.59
Boston Naming	0.61	0.66	0.64	0.70	0.63	0.77	0.62	0.59
<u>SAPS</u>								
Trails A	0.62	0.68	0.67	0.71	0.66	0.79	0.69	0.61
Trails B	0.80	0.83	0.83	0.85	0.84	0.84	0.77	0.81
DSS	0.76	0.79	0.82	0.81	0.78	0.87	0.81	0.75
DSB	0.51	0.50	0.53	0.50	0.49	0.58	0.50	0.45
<u>Fit statistics</u>								
CFI	0.971			0.976		0.972		0.970
TLI	0.958			0.965		0.960		0.956
RMSEA *	0.059 (0.056, 0.062)	0.057 (0.054, 0.061)		0.061 (0.058, 0.065)			0.059 (0.056, 0.063)	
SRMR †	0.035		0.033		0.034			0.037

* RMSEA listed as value (90% confidence interval)

† SRMR calculated from models restricted to complete data

Abbreviations: HS, high school; DWR, delayed word recall; LM, logical memory; SAPS, sustained attention and processing speed; DSS, digit symbol substitution; DSB, digit span backwards; CFI, comparative fit index (>0.90 indicates good fit); TLI, Tucker-Lewis index (>0.90 indicates good fit); RMSEA, root mean squared error of approximation (<0.10 indicates good fit, <0.05 indicates very good fit); SRMR, standardized root mean squared residual (<0.08 indicates adequate fit, <0.05 indicates good fit).