Published in final edited form as:

Epidemiology. 2015 November ; 26(6): 878–887. doi:10.1097/EDE.0000000000000379.

Application of latent variable methods to the study of cognitive decline when tests change over time

Alden L. Gross[1,2], **Melinda C. Power**[1], **Marilyn S. Albert**[3], **Jennifer A. Deal**[1], **Rebecca F. Gottesman**[1,3], **Michael Griswold**[4,5], **Lisa M. Wruck**[6], **Thomas H. Mosley Jr.**[7], **Josef Coresh**[1], **A. Richey Sharrett**[1], and **Karen Bandeen-Roche**[2,5]

[1]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

[2]Johns Hopkins University Center on Aging and Health, Baltimore, MD

[3]Department of Neurology, Johns Hopkins School of Medicine, Baltimore, MD

[4]Center of Biostatistics and Bioinformatics, University of Mississippi Medical Center, Jackson, MS

[5]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

[6]Department of Biostatistics, UNC Gillings School of Global Public Health, Chapel Hill, NC

[7]Department of Medicine, University of Mississippi Medical Center, Jackson, MS

Abstract

Background—The way a construct is measured can differ across cohort study visits, complicating longitudinal comparisons. We demonstrated the use of factor analysis to link differing cognitive test batteries over visits to common metrics representing general cognitive performance, memory, executive functioning, and language.

Methods—We used data from three visits (over 26 years) of the Atherosclerosis Risk in Communities Neurocognitive Study (ARIC-NCS) (N=14,252). We allowed individual tests to contribute information differentially by race, an important factor to consider in cognitive aging. Using generalized estimating equations, we compared associations of diabetes with cognitive change using general and domain-specific factor scores vs. averages of equally weighted standardized test scores.

Results—Factor scores provided stronger associations with diabetes at the expense of greater variability around estimates (e.g., for general cognitive performance, −0.064 SD units/year, SE=0.015, vs −0.041 SD units/year, SE=0.014), which is consistent with the notion that factor scores more explicitly address error in measuring assessed traits than averages of standardized tests.

Conclusions—Factor analysis facilitates use of all available data when measures change over time, and further, it allows objective evaluation and correction for differential item functioning.

Corresponding author: Alden L. Gross, agross14@jhu.edu; Phone: 410-474-3386; Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 2024 E. Monument St., Baltimore, MD 21205.

Keywords

calibration; neuropsychological performance; diabetes; structural equations modeling; cognitive decline; epidemiologic methods; latent variables

Introduction

The study of change in a variable such as cognition requires that it be measured repeatedly and in the same way each time. However, measures often change over time due to refinements in theory and administrative or technical limitations.(1) Solutions are needed that facilitate use of available data to study change and its predictors. One analytic approach to using all available data is to use only tests in common across study visits. This approach potentially discards information from tests used in some but not all visits. Alternatively, scores from all tests might be standardized and averaged together into a composite.(e.g., 2) Although this succeeds in placing cognitive performance across study visits on a common scale (e.g., z-score with mean 0, variance 1), inclusion of equally-weighted cognitive tests that differ across different visits might reflect different constructs over time. As an extreme example, an average of nine memory tests and one speeded test would be entirely different than an average of one memory test and nine speeded tests at another point in time. A third approach, that can use all tests at all visits, is to use latent variable methods, described in the Methods, to derive scores representing latent traits that more adequately represent the same constructs across study visits and that take maximum advantage of existing data.

Latent variable methods have several advantages. First, they allow estimation of cognitive performance on a common metric despite differences in the tests used across assessments.(4–6) Second, they account for varying difficulty and strengths of relationships among tests and take into account only common covariation among tests, thus addressing measurement error.(7) Third, they allow objective evaluation of and adjustment for test-level bias, or differences in test scores by characteristics extraneous to cognitive performance, that may alter the relationship between performance on a given cognitive test and the underlying cognitive trait it represents. In particular, although disparities in cognitive performance by racial/ethnic minority status are substantial,(8–9) and attributable largely to social factors such as educational attainment or accumulation of adverse life experiences,(10–11) differences in measurement properties of specific tests likely also contribute.(12–13) Finally, latent variable methods address incomplete information, as models can be explored that assume item scores not present are missing at random conditional on one's latent status.

This study aimed to demonstrate the application of latent variable methods to cognitive data where tests differ over time. Using data from the Atherosclerosis Risk in Communities Neurocognitive study (ARIC NCS), we derived common factors representing general cognitive performance, executive functioning, memory, and language using all available cognitive data in each study visit. We explored and adjusted for differential item functioning between white and black participants because prior research has documented test bias by race in older adults.(6,11,12) We then contrasted the association of diabetes with cognitive change using the derived cognitive factors vs. corresponding associations based on averages

of standardized tests. We selected diabetes as an exemplar to parallel recent analyses using the same data based on standardized averages of scores.(14) We hypothesized that using psychometrically appropriate methods addresses error in measuring the cognitive traits assessed by the test battery more effectively than taking averages of tests, and thus shows less bias at the expense of more variability around estimates.

Methods

Participants

ARIC recruited N=15,792 adults aged 45 to 64 in 1987–89, representative of four US communities (Washington County, Maryland; suburban Minneapolis, Minnesota; Jackson, Mississippi; and Forsyth County, North Carolina). A primary objective of the ARIC NCS is to evaluate the contribution of vascular risk factors, measured during midlife, to long-term cognitive decline. Three cognitive tests (Digit Symbol Substitution, Delayed Word Recall, and Semantic Fluency) were administered in a 1990–92 visit to all available participants (N=14,252) and again in a 1996–98 visit (N=11,383). Ten tests (Digit Symbol Substitution, Delayed Word Recall, Phonemic Fluency, Logical Memory, Digit Span, incidental learning, Trail Making parts A and B, phonemic fluency, and the Boston naming test) were administered in 2011–13, when participants (N=6,351) were on average 77 years old (range 66, 91). We excluded participants with missing data on diabetes at the 1990–92 visit (N=59). Our analysis includes 14,252 participants tested at least once. The study was approved by institutional review boards at each recruitment site.

Variables

Cognitive performance—Cognitive tests are described in Supplemental materials. The tests administered at each visit are depicted in Figure 1. We took delayed word recall, logical memory, and incidental learning to represent memory; the Trail Making Test, parts A and B and digit symbol substitution to represent executive functioning and speed of information processing; and semantic and phonemic fluency and the Boston Naming tests to represent language. These domains from these and related tests have been empirically derived in the ARIC NCS (15) and elsewhere.(16–18)

Diabetes—Diabetes at the 1990–92 visit was based on self-reported diagnosis by physician, use of medication for diabetes, or fasting blood glucose ≥126 mg/dL.

Control variables—Potential confounders, measured at the 1990–92 visit, included age, sex, indicators for race and ARIC field center (white participants from Minnesota, white participants from Washington County, white participants from Forsyth County, black participants from Forsyth County, black participants from Mississippi), level of education (less than high school, high school or equivalent, more than high school), prevalent coronary heart disease, prevalent stroke, hypertension, total cholesterol, body mass index, cigarette history (never, former, current), apolipoprotein E status (any ε4 allele or none), and drinking history (never, former, current).

Analyses

First, we derived factor scores for general cognitive performance, executive functioning, memory, and language using factor analysis. Factor analysis is a structured approach for describing common covariation among a set of observed indicators, here cognitive tests. We tested and adjusted for differential item functioning, or item-level bias, by race. We then used generalized estimating equations (GEE) to determine the association between diabetes and cognitive change using the derived cognitive factors, adjusting for the covariates and interactions between each covariate and time. We compared estimates using the factors with estimates using equally weighted averages of standardized tests, as described below. Each score was scaled to have a mean of 50 and variance of 10 at the 1990–92 visit. Thus, a score of 50 reflects average cognitive performance at the 1990–92 visit. Scores of +10 and −10 reflect one SD above and below the average relative to the 1990–92 visit.

Derivation of factors for cognitive performance—We estimated a confirmatory factor analysis of the test battery from the 2011–2013 visit. The factor analysis model, also with factor analyses at earlier time points, is presented graphically in Figure 1. The factor analysis corresponds to a 2-parameter logistic item response theory model.(19) These models estimate two sets of parameters for each test. Factor loadings, or weights, describe how well a cognitive test separates persons of low and high ability on the latent trait, or equivalently, how strongly the cognitive test is correlated with other tests in the trait. Thresholds, or boundaries, describe the location on the latent trait where the probability of responding in a given category or better of a test is 50%. In Figure 2, the factor analysis approach is summarized (right panel) and contrasted with the equally weighted average of standardized tests approach (left panel). We examined normalized residuals, calculated using sample and model-estimated polychoric correlations, to evaluate the fit of the model to the data.(20) Normalized residuals are detailed fit statistics for each pairwise correlation among cognitive tests, thus pinpointing specific areas of misfit in the factor analytic model. We transformed raw continuous test scores into as many as 9 categories using an approximately equal-percentile approach while ensuring adequate numbers (greater than N=100) in each category (Supplemental Table 1). This approach allowed us to locate along the latent variable trait specific scores on individual tests where they are found to belong. This is illustrated later in Figure 3. This approach also has advantages of placing tests on a common scale (0 to 9) and accommodates skewed test indicators.(5) Common in cognitive aging research,(21–25) these transformations were based on empirical distributions in the data. In a sensitivity analysis, we derived factors using raw continuously distributed scores and compared results with those for the categorized scores.

Next, we tested for differential item functioning attributable to race. The goal of this analysis is to determine which items appear to be the same by race, and which should be allowed to vary by race. Differential item functioning across groups is present when scores on tests comprising a cognitive factor depend on group membership, controlling for the underlying level of cognitive performance.(26) For example, one item in the Boston Naming Test, a test of language, is a picture of a volcano. Individuals who may have never seen a picture of a volcano may fail to identify it regardless of their language ability. If the Boston Naming Test showed differential functioning in favor of whites, then white participants of the same

underlying level of cognitive ability level as black participants (as indicated by performance across tests) would be expected to score higher on that test than black participants on average. We used multiple-group factor analysis to test for differential item functioning by race.(27) The model is comprised of a series of probit regression relationships linking each cognitive test to the underlying cognitive latent variable. A difference in factor loadings between groups suggests the strength of the common correlation with a test differs by group. A difference in thresholds, controlling for performance on other tests, suggests a test is systematically more difficult in one group than another at a given level of cognitive performance. Conceptually, differential item functioning in factor loadings and thresholds corresponds to tests of effect modification and confounding, respectively. We empirically identified cognitive tests with no differential item functioning, or anchors, using a model for each test.(28) Specifically, we identified anchor items using iterative likelihood ratio tests to test the equivalence of factor loadings and thresholds for non-anchor tests across groups by freeing one loading or threshold between groups at a time as dictated by model modification indices.(29)

After identifying and adjusting for differential item functioning, we derived factor scores for general cognitive performance, memory, executive functioning, and language, with a factor at each of the three study visits. To ensure cognitive performance was measured on the same metric in the early study visits, when only 3 tests were administered, as they were in 2011–13 when they were part of the 10-test battery, loadings and thresholds for tests measured at multiple visits were fixed to be equal across visit. By constraining test thresholds to be the same across visit, change in cognitive performance over time is reflected in the levels of the latent variables at each study visit, which are then estimated as factor scores. We acknowledge here the assumption that tests change at a similar rate; such longitudinal invariance has been demonstrated in several other studies using similar cognitive test batteries.(16,17) This assumption was necessary given the structural missingness across study visits. We repeated this procedure for the domains: executive functioning/processing speed (using Digit Symbol, Trail Making Test parts A and B at the 2011–13 visit), memory (using delayed word recall, logical memory, incidental learning at the 2011–13 visit), and language (using phonemic and semantic fluency, Boston Naming at the 2011–13 visit). Domain-specific factors at the first two visits are informed entirely by the single cognitive test items measured at that visit. The loading and threshold of the items, defined from the factor analysis at the 2011–13 visit due to model constraints, determines the scale of the factor and causes the scale to be the same across study visits.

Factor scores were estimated using the regression-based method and scaled to have a mean of 50 and standard deviation of 10 at the 1990–92 visit.(30) We used Mplus software to estimate models, using a maximum likelihood estimator with robust standard error estimation under the Expectation-maximization algorithm.(31) Mplus syntax for factor analyses unadjusted and adjusted for differential item functioning are provided in Supplemental Information 2 and 3.

We used a “test information” plot to quantify reliability with which general and domain-specific factors were measured over the range of cognitive performance.(32) Reliability is based on the relationship between observed test scores and the estimated latent factor.(32)

Reliability, calculated as the complement of the square of the standard error of measurement of each observation, can vary over the range of cognitive performance. The necessary standard of reliability depends on the intended purpose of an instrument; Nunnally (33) recommends minimum reliabilities of 0.80 and 0.90 as rules of thumb for between-persons analysis and within-persons analysis, respectively.

Equally weighted averages of standardized tests—To provide sample-average composite scores comparable to the factor scores, we standardized each averaged composite to have a mean of 50 and SD of 10 at the 1990–92 visit. We standardized each of the 10 cognitive tests at the 2011–13 visit, and scaled scores at other visits to that visit. We then took an average of the three test scores at the 1990–92 visit and the 1996–98 visit and the 10 tests at the 2011–2013 visit to construct an equally weighted composite average of cognitive tests. We did the same for cognitive domains. For executive functioning, we used Digit Symbol Substitution at the 1990–92 and 1996–98 visits and Digit Symbol Substitution and Trails A and B at the 2011–13 visit. The average score for memory used delayed word recall at the 1990–92 and 1996–98 visits and delayed word recall, Logical Memory, and incidental learning at the 2011–13 visit. The average score for language used phonemic fluency at the 1990–92 and 1996–98 visits and semantic and phonemic fluency and Boston Naming at the 2011–13 visit.

Comparison of the association of diabetes with cognitive change across cognitive measures—We used GEE methods to estimate associations between diabetes and change in each cognitive outcome, parameterized using either factor scores or the average of individual cognitive test scores.(34) We evaluated the approaches by comparing the magnitude of associations and standard errors for the effect of diabetes on change. The timescale was time since the 1990–92 study visit. Primary independent variables were study visit, diabetes status at the 1990–92 visit, and interactions between diabetes and time since the 1990–92 visit. We adjusted for covariates, listed earlier, and their interactions with time. GEE models used an unstructured correlation matrix with a Huber-White robust variance estimator (35) and were estimated using Stata 13.1.(36)

Sensitivity analyses—In addition to GEE methods using factor scores, we estimated the association of diabetes with cognitive decline simultaneously with the factor analysis in a structural equations model specifying a latent growth curve across the factors (Supplemental Figure 1)(37). In another sensitivity analysis, we compared associations between diabetes and general cognitive performance based on all available tests to associations based on a model using only the three tests administered at all study visits.

Missing data handling—Cognitive test scores were missing less than 5% of the time except for Trails B at the 2011–13 study visit (13% missing) (Supplemental Table 2). Taking equally weighted averages of standardized tests uses only complete cases, which assumes a missing completely at random mechanism in the data. Missingness in factor analysis models was handled using a maximum likelihood estimator with robust standard error estimation. This approach makes a less restrictive missing data assumption than the standardize and average approach by assuming missingness in specific cognitive tests are missing at random

conditional on other cognitive tests in the model. This is a reasonable assumption because structural missingness in individual tests at earlier ARIC visits is attributable to study design, not to participants' cognitive ability.

Results

Characteristics of the study sample are in Table 1. At baseline in 1990–92, the sample was mostly female (55%), white (75%), and had at least a high school education (78%). The mean age was 57 years in 1990–92 (range 46, 70 years) and 76 years in 2011–2013 (range 67, 91).

Factor analysis

Factor analysis with categorical indicators allowed us to empirically place thresholds where they belong along the latent trait, as shown in Figure 3 for general cognitive performance at the 2011–13 visit before adjustment for differential item functioning by race. Based on normalized residuals (e.g., flagging normalized residuals with absolute values > 2.0), in the model of general cognitive performance we allowed residual correlations between delayed word recall (DWR), logical memory (LM), incidental learning (INCLRN), and Trail Making (TMT), part A. Tests fit well in the resulting model (Supplemental Table 3). The broad spread of boundary response thresholds for tests over the range of cognitive performance, shown by vertical bars in Figure 3, suggests a wide dynamic range of measurement consistent with a factor optimized for longitudinal analysis across a range of ability. From Figure 3, the model rates generation of at least 25 words on semantic fluency as comparable in difficulty to recalling 8 words on delayed word recall. Further, generating ≤14 words on semantic fluency (AN) is rated as more impairment than recalling ≤2 digit-symbol pairs on the incidental learning test.

We tested for differential item functioning by race (Table 2). Using all other tests as tentative indicators, initial modeling identified delayed word recall and logical memory as being free of differential item functioning.(38) The remaining tests were evaluated for differential item functioning. Models identified differential item functioning in factor loadings for Trails A, Digit Symbol, Boston Naming, phonemic fluency, and digit span backwards; these measures were all more highly inter-correlated among black participants than white participants. Models also detected differential item functioning in thresholds for Trails A and B, digit span backwards, and Boston Naming such that these tests were more difficult for black participants, controlling for general cognitive performance. Models detected no differential item functioning in incidental learning or semantic fluency after adjustment for differential item functioning in other tests. Memory and language factor models showed no differential item functioning. The executive functioning factor demonstrated differential item functioning in factor loadings for Trail Making, part A.

Differential item functioning-adjusted item factor loadings for the general and domain-specific factor models are shown in Table 2. For the general factor, all loadings were ≥0.47, suggesting the factor represents general performance and not a particular domain. All loadings were also high for domain-specific factor models.

Precision of the latent factor measurement

Figure 4 shows the reliability of the measurement of general and domain-specific factors for each study visit. As expected given the expanded battery, general cognitive performance for the 2011–13 visit was precise (>0.80) across a 4 SD range in the middle of the distribution which encompassed approximately 67% of the sample. Reliability at the earlier visits fell between 0.65 and 0.75 across the range of cognitive performance. Reliability for memory, language, and executive functioning domains were lower, as expected since they include fewer cognitive tests.

Comparison across cognitive measures of the association of diabetes with cognitive change

GEE models (Table 3) all fit the data well, as indicated by inspecting residuals. Coefficients in table 3 represent the difference in cognitive slope per decade in SD units by diabetes status; results are shown for factors with and without adjustment for differential item functioning. The group with diabetes demonstrated a steeper average decline in the general cognitive performance factor of −0.058 SD units per decade (SE=0.016) over the entire study period (differential item functioning-adjusted, −0.064 SD units per decade, SE=0.015), compared to −0.041 SD units per decade (SE=0.014) using the average of three standardized tests at the 1990–92 and 1996–98 visits and 10 tests at the 2011–13 visit. Standard errors of the estimate were larger than those using averages of standardized scores.

For decline in memory and language, magnitudes of overall associations with diabetes were similarly stronger using factors than using averages of standardized scores, and standard errors were larger (Table 3). For decline in executive functioning, diabetes was more strongly associated with the average score than the differential item functioning-adjusted factor, but not with the non-differential item functioning-adjusted factor. Consistent with previous findings,(14) the association with diabetes was stronger for executive functioning than for memory.

Sensitivity analyses

Associations between diabetes and general cognitive performance were comparable when examined directly from latent growth models that simultaneously estimated measurement models (B=−0.032 SD, SE=0.006), using a general factor score based only on the three common tests across visits (B=−0.046, SE=0.013), and using continuously distributed tests in the factor analysis (B=−0.077, SE=0.017). Factor analysis using continuous indicators fit poorly to the data (Root Mean Square Error of Approximation=0.24; Comparative Fit Index=0.00), but its use did not change inferences with diabetes. Corresponding overall fit indices for the categorical case using weighted least squares estimation was much better (Root Mean Square Error of Approximation=0.08; Comparative Fit Index=0.977).

Discussion

We calibrated general and domain-specific cognitive performance across study visits in which different but overlapping cognitive tests were administered at each visit. Associations of diabetes with cognitive change were generally stronger using factor scores than with

corresponding equally weighted averages of standardized tests, at the expense of greater variability around estimates. This is consistent with the hypothesis that latent variables account for error in measuring cognitive traits better than averages of test scores, and thus more accurately depict the relationship with diabetes. This manifests in three primary ways. First, scores predicting the latent variables give greater weight to tests more correlated with other tests in the cognitive domain than investigator-assigned weighting (e.g., equal weighting). These weights are the loading factors in Figure 3. Second, the latent variable model provides an appropriate location for scores on individual tests (thresholds in Figure 3). By contrast, the average of standardized scores approach forces a z-score of −1.5 SD units on one test to represent an equivalent level of cognitive impairment to a z-score of −1.5 SD units on another test, which may not be appropriate. Third, the differential item functioning adjustment may provide a better measure of the trait affected by diabetes in each racial group by tying measured performance on cognitive measures to the better-measured underlying cognitive ability, thus endeavoring to filter extraneous characteristics related to race.

Despite the added analytic complexity, factor analysis is appropriate when test batteries change over time because it facilitates use of all available data appropriately. In ARIC NCS, factor analysis combines the full test battery at the 2011–2013 visit with the limited test battery in the earlier visits in two primary ways. First, it appropriately weights items as explained above for the general factor, which is represented by only three tests at the earlier visits. Second, for both the general factor and the individual domains, in which only one test is available at earlier visits, factor analysis determines an appropriate placement of thresholds for specific test scores, as explained above. The factor analysis approach we used is extendable to other studies in which measures of a construct differ across time; in fact, *time* is not the only dimension for which the approach could be used. Factor analysis has been successfully applied to harmonize cognitive performance,(5,39–40) physical functioning,(41) and depressive symptoms (42) across *datasets*.

Strengths of the study are the well-characterized ARIC NCS sample and carefully collected longitudinal data over up to 26 years. We used contemporary latent variable methods to measure cognitive performance permitting use of all available cognitive data and accounting for differential item functioning by race. In contrast, simple averages of different numbers and types of standardized tests across visits could reflect different constructs at each visit, complicating analyses of within-person change.

Several caveats are appropriate. First, in testing differential item functioning, we used empirical criteria for identifying cognitive tests free of differential functioning. Cognitive tests without differential item functioning by race, or anchor tests, are necessary against which to test differential functioning in other tests, but results may not be the same in different settings. Differential item functioning in the same direction and of similar magnitudes for all tests cannot be detected mathematically; such differences would translate into overall group differences. Fortunately, our findings were similar when we used scores not adjusted for differential item functioning. A second limitation is that factor scores we used in the primary analysis can potentially, but not always, provide biased estimates of an underlying trait (43); we used them because they are more convenient for investigators less

familiar with factor analysis and can be distributed widely through data distributions. We note that we obtained similar inferences regarding the association of diabetes with cognitive change using a direct approach in which we entered diabetes into a latent growth model of the latent variables. Another study limitation is that the factor structure based on empirical correlations among tests may vary somewhat depending on characteristics of a sample, such as by dementia status or language ability.(44) When this possibility has been tested in data, it is rarely supported.(16–18,45–48) Thus, we believe factor analysis in ARIC NCS is appropriate across the range of cognitive performance we studied. We also used categorical transformations of cognitive tests in factor analyses. Although categorization may coarsen the data, it diminishes the influence of truncation and outliers.(5) A final limitation is that temporal harmonization provided by factor analysis relies on common tests across time; the approach is not applicable in settings without anchor tests or items unless further assumptions are made.

Optimization of cognitive measures is important. We harmonized available cognitive data into factors representing general cognitive performance, executive functioning, memory, and language in ARIC NCS. We corrected the factors for differential functioning by race. Factors provided stronger estimates of associations with diabetes compared to averages of standardized tests. Latent variable approaches may be useful other studies with differing cognitive measures across visits.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the staff and participants of the ARIC study for their important contributions.

Funding/Support: ARIC National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). Neurocognitive (ARIC-NCS) data is collected by U01 HL096812, HL096814, HL096899, HL096902, HL096917 with previous brain MRI examinations funded by R01-HL70825. Dr. Gross was supported by grant R03AG045494 from the National Institute on Aging. Dr. Gottesman was supported by grant R01AG040282 from the National Institute on Aging. Dr. Power was supported by T32AG027668 from the National Institute on Aging. Dr. Bandeen-Roche was supported by P50AG005146 from the National Institute on Aging.

ARIC-NCS Steering Committee: Thomas Mosley (Chair), Josef Coresh (Co-Chair), Marilyn Albert, Alvaro Alonso, Christie Ballantyne, Eric Boerwinkle, David Couper, Gerardo Heiss, Clifford Jack, Barbara Klein, Ronald Klein, David Knopman, Natalie Kurinij (NEI Project Officer), Claudia Moy (NINDS Project Officer), and Jacqueline Wright (NHLBI Project Officer). Ex Officio Members: Laura Coker, Aaron Folsom, Rebecca Gottesman, Richey Sharrett, Lynne Wagenknecht, and Lisa Miller Wruck.

ARIC-NCS Data Analysis Committee: Richey Sharrett (Chair), Karen Bandeen-Roche (Senior Statistician), Andrea Christman, Joe Coresh, Jennifer Deal, Rebecca Gottesman, Michael Griswold, Alden Gross, Thomas Mosley, Melinda Power, Andreea Rawlings, and Lisa Wruck. Shoshana Ballew (Epidemiologist coordinator).

ARIC-NCS Neurocognitive Committee: Thomas Mosley (Chair), Rebecca Gottesman (Co-Chair), Alvaro Alonso, Laura Coker, David Couper, David Knopman, Guy McKhann, Ola Selnes, and Richey Sharrett.

Abbreviations

ARIC NCS	Atherosclerosis Risk in Communities Neurocognitive study
SD	Standard deviation
GEE	Generalized estimating equations

References

1. McArdle JJ, Grimm KJ, Hamagami F, Bowles RP, Meredith W. Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. Psychol Methods. 2009; 14(2):126–149. [PubMed: 19485625]
2. Willis SL, Tennstedt SL, Marsiske M, Ball K, Elias J, Koepke KM, Wright E. ACTIVE Study Group. Long-term effects of cognitive training on everyday functional outcomes in older adults. JAMA. 2006; 296:2805–2814. [PubMed: 17179457]
3. Wilson RS, Beckett LA, Barnes LL, Schneider JA, Bach J, Evans DA, Bennett DA. Individual differences in rates of change in cognitive abilities of older persons. Psychol Aging. 2002; 17(2): 179–193. [PubMed: 12061405]
4. Crane PK, Narasimhalu K, Gibbons LE, Mungas DM, Haneuse S, Larson EB, van Belle G. Item response theory facilitated calibrating cognitive tests and reduced bias in estimated rates of decline. J Clin Epidemiol. 2008; 61:1018–1027. [PubMed: 18455909]
5. Gross AL, Sherva R, Mukherjee S, Newhouse S, Kauwe JSK, Munsie LM, Crane PK. for the Alzheimer's Disease Neuroimaging Initiative, GENAROAD Consortium, and AD Genetics Consortium. Calibrating longitudinal cognition in Alzheimer's disease across diverse test batteries and datasets. Neuroepidemiology. 2014; 43(3–4):194–205. [PubMed: 25402421]
6. Mungas DM, Reed BR, Kramer JH. Psychometrically matched measures of global cognition, memory, and executive function for assessment of cognitive decline in older persons. Neuropsychology. 2003; 17:380–392. [PubMed: 12959504]
7. Balsis S, Unger AA, Benge JF, Geraci L, Doody RS. Gaining precision on the Alzheimer's Disease Assessment Scale-cognitive: a comparison of item response theory-based scores and total scores. Alzheimers Dement. 2012; 8(4):288–294. [PubMed: 22465173]
8. Schneider ALC, Sharrett AR, Gottesman RF, Coresh J, Coker L, Wruck L, Mosley TH. Normative data for 8 neuropsychological tests in older blacks and whites from the Atherosclerosis Risk in Communities (ARIC) Study. Alzheimer Dis Assoc Disord. 2014 Epub ahead of print.
9. Schwartz BS, Glass TA, Bolla KI, Stewart WF, Glass G, Rasmussen M, Bandeen-Roche K. Disparities in cognitive functioning by race/ethnicity in the Baltimore Memory Study. Environ Health Perspect. 2004; 112(3):314–320. [PubMed: 14998746]
10. Sisco S, Gross AL, Shih RA, Sachs BC, Glymour MM, Bangen KJ, Manly JJ. The role of early life educational quality and literacy in explaining racial disparities in cognition in late life. J Gerontol B Psychol Sci Soc Sci. 2014 [Epub ahead of print].
11. Zsembik BA, Peek MK. Race differences in cognitive functioning among older adults. J Gerontol B Psychol Sci Soc Sci. 2001; 56(5):S266–S274. [PubMed: 11522808]
12. Jones RN. Racial bias in the assessment of cognitive functioning of older adults. Aging and Mental Health. 2003; 7(2):83–102. [PubMed: 12745387]
13. Teresi JA, Kleinman M, Ocepek-Welikson K. Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. Stat Med. 2000; 19(11–12):1651–1683. [PubMed: 10844726]
14. Rawlings AM, Sharrett AR, Schneider AL, Coresh J, Albert M, Couper D, Selvin E. Diabetes in midlife and cognitive change over 20 years: a cohort study. Ann Intern Med. 2014; 161(11):785–793. [PubMed: 25437406]
15. Rawlings AM, Bandeen-Roche K, Carlson MC, Coker LH, Gottesman RF, Mosley TH, Penman AD, Selnes OA, Sharret AR. Cognitive domains in elderly blacks and whites in the Atherosclerosis Risk in Communities Neurocognitive Study. (in preparation).

16. Hayden KM, Jones RN, Zimmer C, Plassman BL, Browndyke JN, Pieper C, 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 Disease and Associated Disorders. 2011; 25(2):128–137. [PubMed: 21606904]
17. Park LQ, Gross AL, McLaren DG, Pa J, Johnson JK, Mitchell M, Manly JJ. Alzheimer's Disease Neuroimaging Initiative. Confirmatory factor analysis of the ADNI neuropsychological battery. Brain Imaging and Behavior. 2012; 6(4):528–539. [PubMed: 22777078]
18. Siedlecki KL, Manly JJ, Brickman AM, Schupf N, Tang MX, Stern Y. Do neuropsychological tests have the same meaning in Spanish speakers as they do in English speakers? Neuropsychology. 2010; 24(3):402–411. [PubMed: 20438217]
19. Takane Y, de Leeuw J. On the relationship between item response theory and factor analysis of discretized variables. Psychometrika. 1987; 52:393–408.
20. Bollen, KA. Structural equations with latent variables. Wiley-Interscience; 1989.
21. Crane PK, Carle A, Gibbons LE, Insel P, Mackin RS, Gross A, Jones RN, Mukherjee S, Curtis SM, Harvey D, Weiner M, Mungas D. Alzheimer's Disease Neuroimaging Initiative. Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). Brain Imaging Behav. 2012; 6(4):502–516. [PubMed: 22782295]
22. Gibbons LE, Carle AC, Mackin RS, Harvey D, Mukherjee S, Insel P, Curtis SM, Mungas D, Crane PK. Alzheimer's Disease Neuroimaging Initiative. A composite score for executive functioning, validated in Alzheimer's Disease Neuroimaging Initiative (ADNI) participants with baseline mild cognitive impairment. Brain Imaging Behav. 2012; 6(4):517–527. [PubMed: 22644789]
23. Gross AL, Jones RN, Fong TG, Tommet D, Inouye SK. Calibration and validation of an innovative approach for estimating general cognitive performance. Neuroepidemiology. 2014; 42:144–153. [PubMed: 24481241]
24. Jones RN, Rudolph JL, Inouye SK, Yang FM, Fong TG, Milberg WP, Tommet D, Metzger ED, Cupples LA, Marcantonio ER. J Clin Exp Neuropsychol. 2010; 32(10):1041–1049. [PubMed: 20446144]
25. Mungas D, Reed BR, Marshall SC, González HM. Development of psychometrically matched English and Spanish language neuropsychological tests for older persons. Neuropsychology. 2000; 14(2):209–223. [PubMed: 10791861]
26. Holland, PW.; Wainer, H. Differential item functioning. Hillsdale. NJ: Lawrence Erlbaum; 1993.
27. Jöreskog KG, Goldberger AS. Estimation of a model with multiple indicators and multiple causes of a single latent variable. Journal of the American Statistical Association. 1975; 70(351):631–639.
28. Woods CM. Empirical Selection of Anchors for Tests of Differential Item Functioning. Applied Psychological Measurement. 2009; 33:42–57.
29. Steiger JH. Structural model evaluation and modification: An interval estimation approach. Multivariate Behavioral Research. 1990; 25(2):173–180. [PubMed: 26794479]
30. Asparouhov, T.; Muthén, B. [Accessed July 10, 2014] Plausible values for latent variables using Mplus. Technical Report. 2010. from http://www.statmodel.com/download/Plausible.pdf
31. Muthen, LK.; Muthen, BO. Mplus user's guide: Seventh Edition. Los Angeles, CA: Muthen & Muthen; 1998–2012.
32. Green BF, Bock RD, Humphreys LG, Linn RL, Reckase MD. Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement. 1984; 21(4):347–360.
33. Nunnally, JC. Psychometric theory. 2nd. New York: McGraw-Hill; 1978.
34. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73(1):13–22.
35. Huber, PJ. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. Berkeley, CA: University of California Press; 1967. The behavior of maximum likelihood estimates under nonstandard conditions; p. 221-233.
36. StataCorp. Stata Statistical Software: Release 13. College Station, TX: StataCorp LP; 2013.
37. Johnson JK, Gross AL, Pa J, McLaren DG, Park LQ, Manly JJ. for the Alzheimer's Disease Neuroimaging Initiative. Longitudinal change in neuropsychological performance using latent growth models: a study of mild cognitive impairment. Brain Imaging and Behavior. 2012; 6(4): 540–550. [PubMed: 22562439]

38. Woods CM, Oltmanns TF, Turkheimer E. Illustration of MIMIC-Model DIF Testing with the Schedule for Nonadaptive and Adaptive Personality. J Psychopathol Behav Assess. 2009; 31(4): 320–330. [PubMed: 20442793]

39. Gross AL, Jones RN, Fong TG, Tommet D, Inouye SK. Calibration and validation of an innovative approach for estimating general cognitive performance. Neuroepidemiology. 2014; 42:144–153. [PubMed: 24481241]

40. Crane PK, Carle A, Gibbons LE, Insel P, Mackin RS, Gross AL, Mungas D. Alzheimer's Disease Neuroimaging Initiative. Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). Brain Imaging Behav. 2012; 6(4):502–516. [PubMed: 22782295]

41. Gross AL, Jones RN, Inouye SK. Development of an expanded measure of physical functioning for older persons in epidemiologic research. Research on Aging. 2014 [Epub ahead of print].

42. Wahl I, Löwe B, Bjorner JB, Fischer F, Langs G, Voderholzer U, Rose M. Standardization of depression measurement: a common metric was developed for 11 self-report depression measures. J Clin Epidemiol. 2014; 67(1):73–86. [PubMed: 24262771]

43. Estabrook R, Neale M. A Comparison of Factor Score Estimation Methods in the Presence of Missing Data: Reliability and an Application to Nicotine Dependence. Multivariate Behav Res. 2013; 48(1):1–27. [PubMed: 24049215]

44. Delis DC, Jacobson M, Bondi MW, Hamilton JM, Salmon DP. The myth of testing construct validity using factor analysis or correlations with normal or mixed clinical populations: lessons from memory assessment. J Int Neuropsychol Soc. 2003; 9(6):936–946. [PubMed: 14632252]

45. 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. [PubMed: 21038965]

46. Mungas DM, Widaman KF, Reed BR, Tomaszewski SF. Measurement invariance of neuropsychological tests in diverse older persons. Neuropsychology. 2011; 25(2):260–269. [PubMed: 21381830]

47. 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. [PubMed: 18444718]

48. Tuokko HA, Chou PH, Bowden SC, Simard M, Ska B, Crossley M. Partial measurement equivalence of French and English versions of the Canadian Study of Health and Aging neuropsychological battery. Journal of the International Neuropsychological Society. 2009; 15(3): 416–425. [PubMed: 19402928]

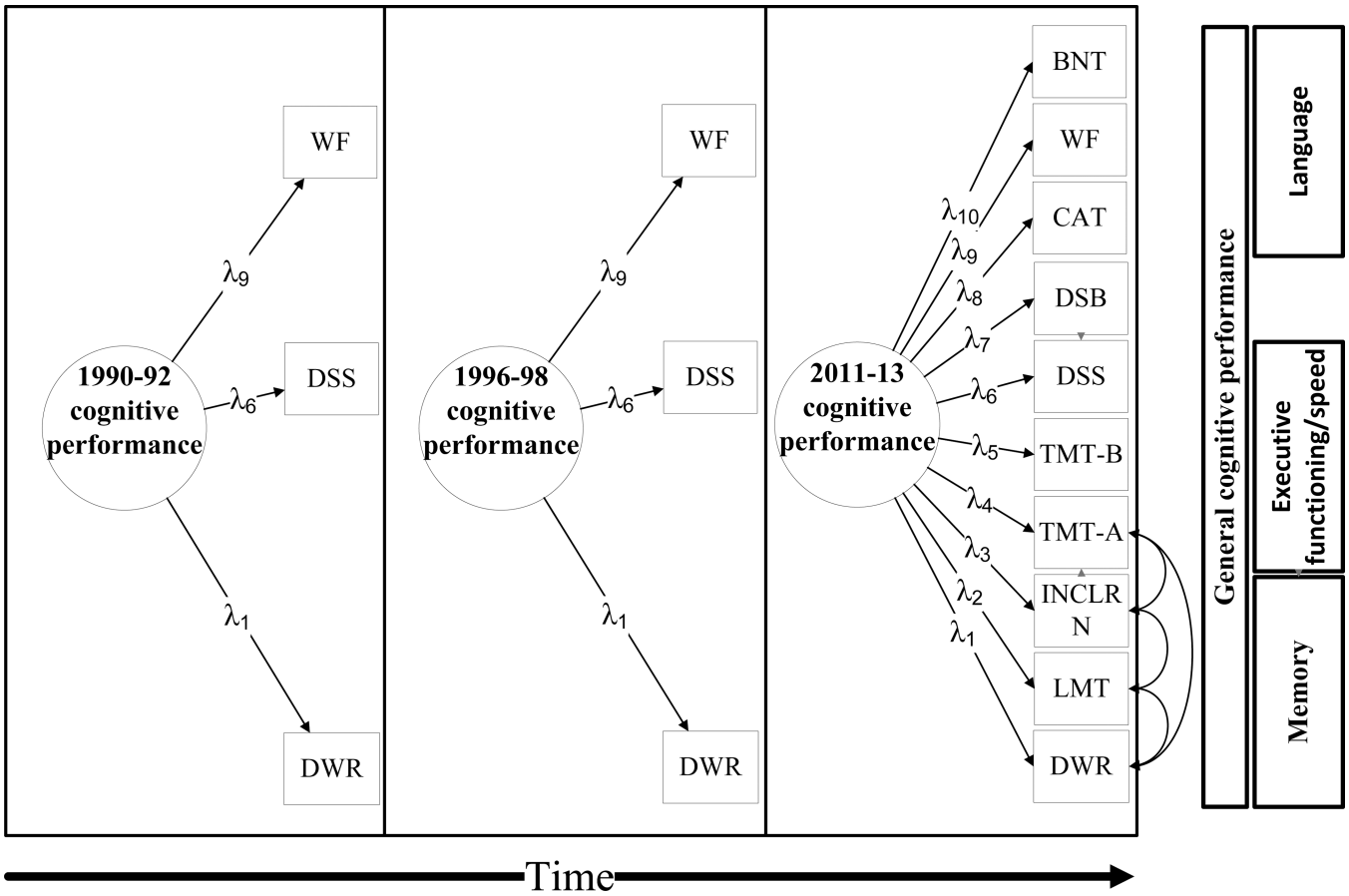

Figure 1. Graphical representation of cognitive tests available at each ARIC study visit

This figure is a structural equations model representing the final longitudinal model described in the Methods. Separate models were estimated for general cognitive performance, memory, executive functioning, and language; see Methods for details. WF: Word (phonemic) fluency (count of words recalled); BNT: Boston Naming Test (number of correct responses); AN: Animal (semantic) fluency (count of words recalled); DSB: Digit span backwards (sum of two trials of the maximum span); DSS: Digit symbol substitution (number of correct digit symbol pairs); TMT: Trail Making Test (seconds to complete); INCLRN: Incidental learning (number of correct digit symbol pairs recalled); LM: Logical memory (sum of recall for 2 stories); DWR: delayed word recall (sum of words recalled from one trial). Curved arrows between TMT-A, INCLRN, LMT, and DWR represent correlations between these items added due to analysis of normalized residuals (see Results).

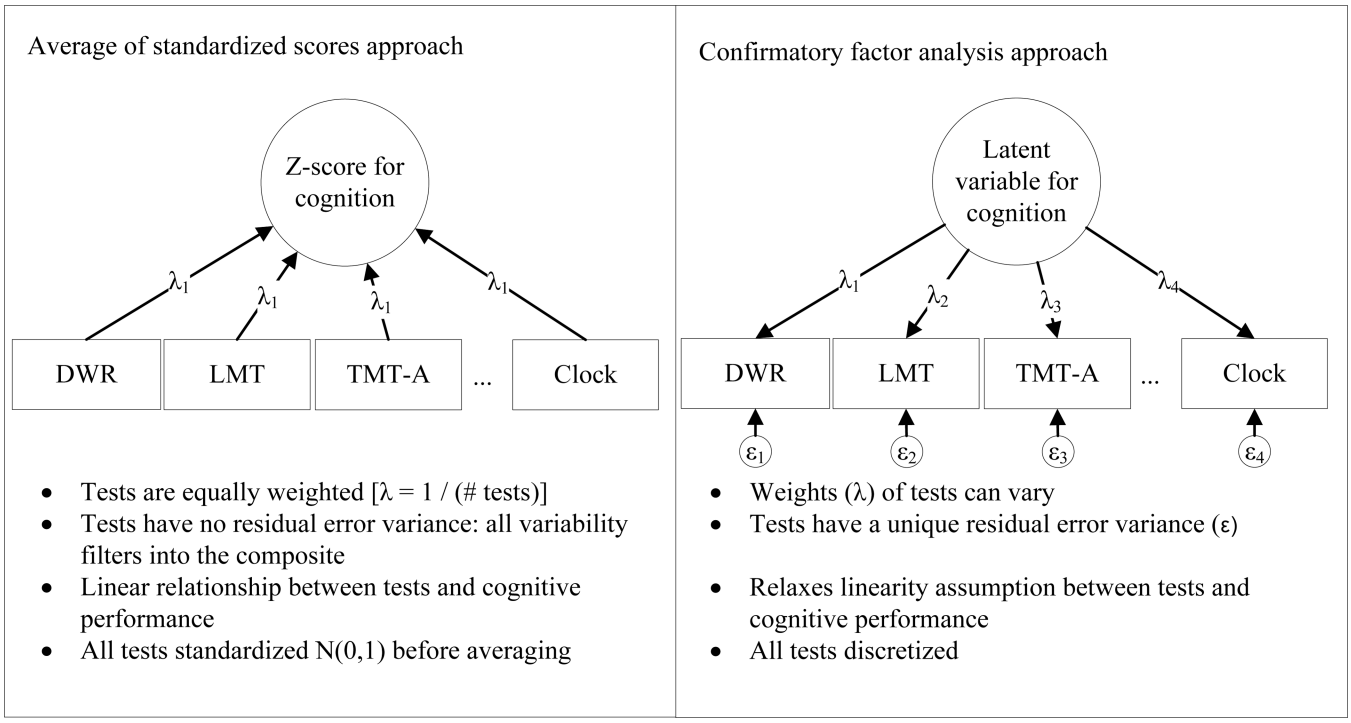

Figure 2. Comparison of Two Approaches for Deriving Summary Scores from a Neuropsychological Test Battery

Measurement models for two approaches, averaging standardized versions of tests and single-factor analysis with categorical variables, are contrasted. Cognitive test scores are provided as examples; refer to Methods for the full neuropsychological test battery.

DWR: delayed word recall; LMT: Logical Memory Test; TMT-A: Trail Making Test, Part A.

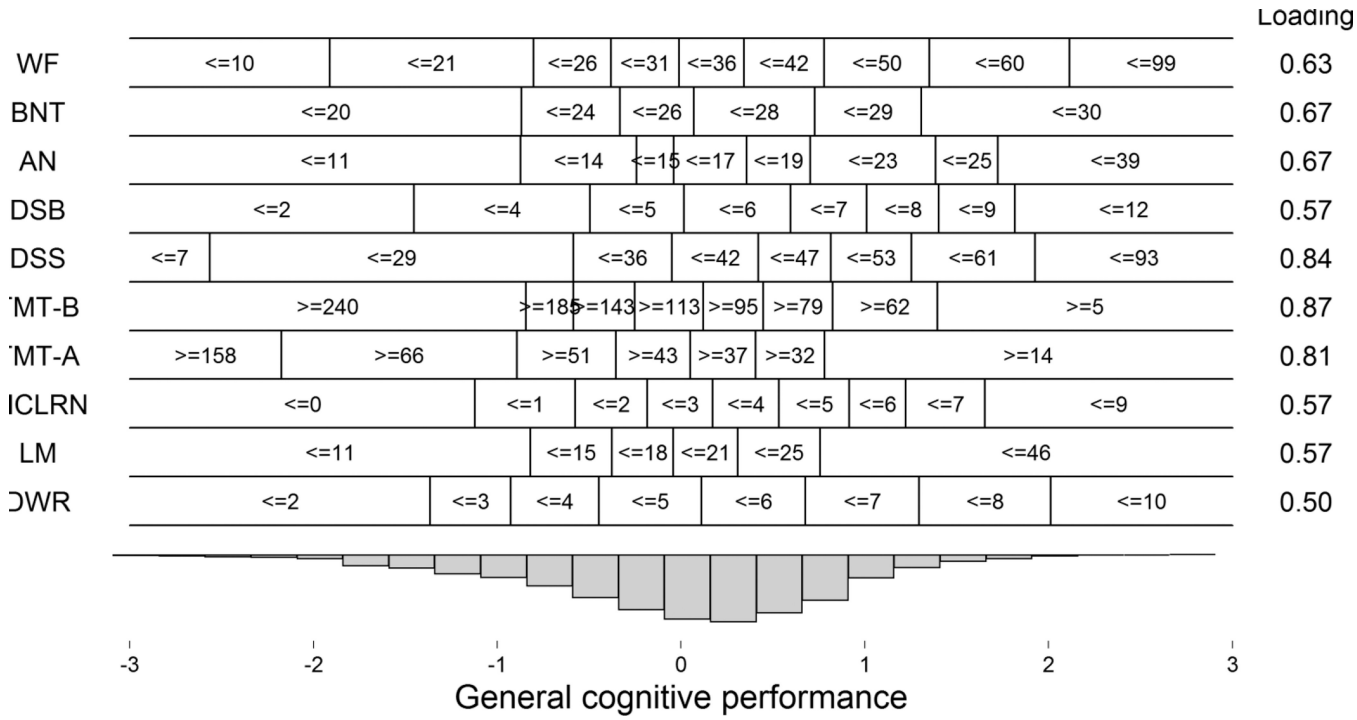

Figure 3. Item loadings and thresholds from the factor analyses for general cognitive performance

Graphical representation of results from the general cognitive performance factor analysis at the 2011–13 ARIC NCS visit. Factor loadings at right represent correlations between a test and the latent variable. Thresholds for responses to each test on the latent variable are shown by vertical boundaries in Appendix Figure 1, and denote the location along the latent variable of general cognitive performance (x axis) where tests provide optimal measurement precision. A histogram of the estimated general cognitive performance factor score in the 2011–13 participant sample, derived from the model that estimated these thresholds, is shown at the bottom. Some parameters (for TMT-A, DSS, BNT, WF, DSB) were estimated separately by race group to account for differential item functioning (see Table 2 and Methods).

WF: Word (phonemic) fluency (count of words recalled); BNT: Boston Naming Test (number of correct responses); AN: Animal (semantic) fluency (count of words recalled); DSB: Digit span backwards (sum of two trials of the maximum span); DSS: Digit symbol substitution (number of correct digit symbol pairs); TMT: Trail Making Test (seconds to complete); INCLRN: Incidental learning (number of correct digit symbol pairs recalled); LM: Logical memory (sum of recall for 2 stories); DWR: delayed word recall (sum of words recalled from one trial).

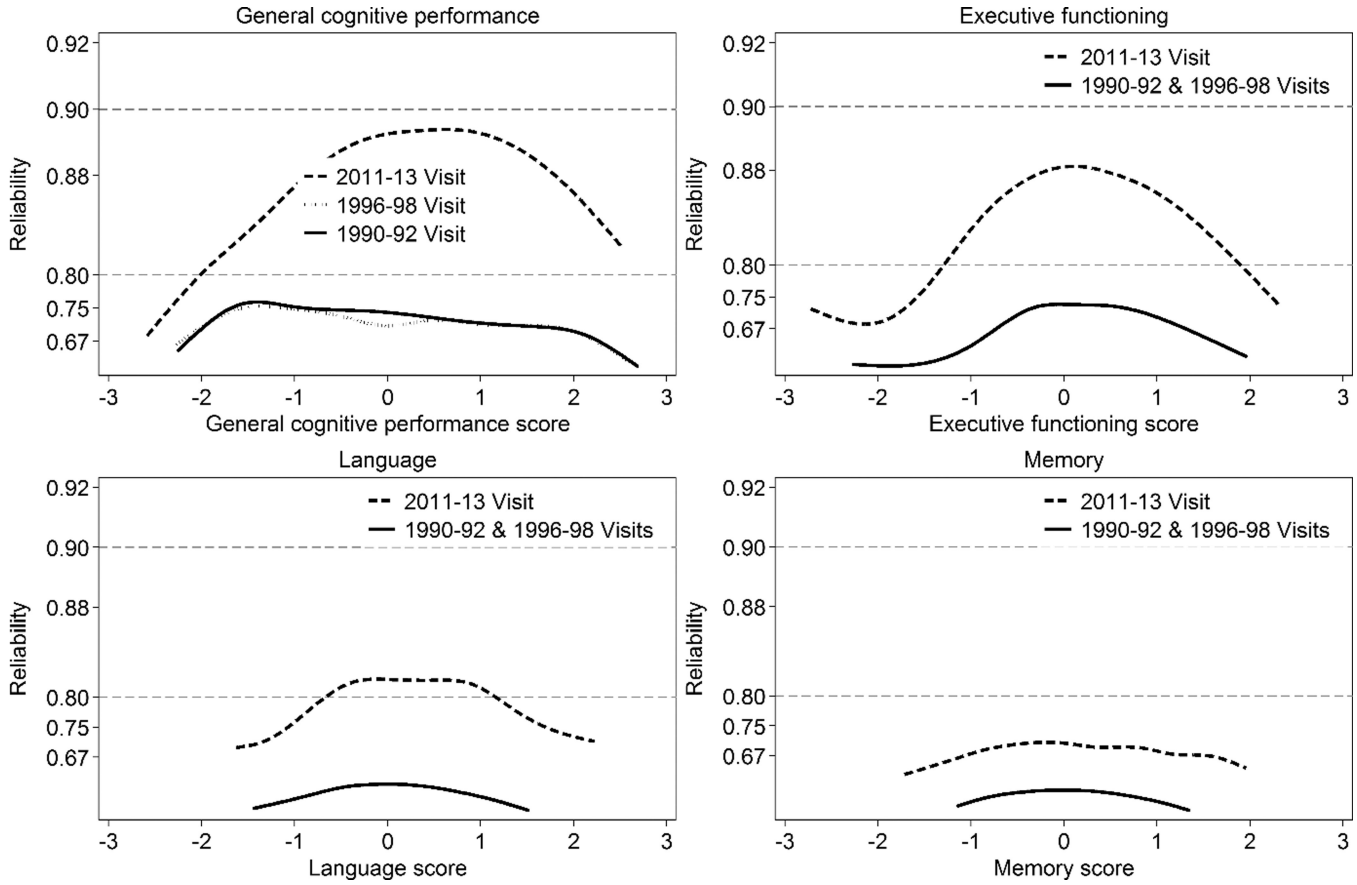

Figure 4. Precision of General and Domain-Specific Cognitive Factors as a Function of Performance Level: Results from ARIC NCS (N=14,252)

Reliability for the general (top left), executive functioning (top right), language (bottom left), and memory (bottom right) factors are plotted over the range of their respective values. Plots for the first two visits overlap almost completely because the same tests were used. Horizontal dashed lines at reliabilities of 0.90 and 0.80 indicate acceptable reliability for within-persons analysis and between-persons analysis, respectively (33). Reliability = 1 – 1 / Information.

Table 1

Baseline characteristics of the ARIC NCS (N=14,252)

Characteristic	Full sample (N=14252)	Diabetes (N=2152)	No diabetes (N=12100)
	Mean (SD) or n (%)	Mean (SD) or n (%)	Mean (SD) or n (%)
Age, mean (SD)	57.0 (6)	58.2 (6)	56.8 (6)
Sex (female), n (%)	7891 (55)	1149 (53)	6742 (56)
Race (White), n (%)	10742 (75)	1264 (59)	9478 (78)
Education, n (%)			
Less than high school	3110 (22)	726 (34)	2384 (20)
High school	5898 (41)	837 (39)	5061 (42)
More than high school	5224 (37)	586 (27)	4638 (38)
Field center, n (%)			
Forsyth County, NC	3669 (26)	448 (21)	3221 (27)
Jackson, MS	3083 (22)	794 (37)	2289 (19)
Mineapolis, MN	3817 (27)	377 (18)	3440 (28)
Washington County, MD	3683 (26)	533 (25)	3150 (26)
Stroke, n (%)	269 (2)	96 (5)	173 (1)
Hypertension, n (%)	5117 (36)	1264 (59)	3853 (32)
Total cholesterol (mmol/L), mean (SD)	5.4 (1)	5.5 (1)	5.4 (1)
Body mass index (kg/m^2, mean (SD)	28.0 (5)	31.2 (6)	27.4 (5)
Smoking status (n (%))			
Never	5665 (40)	896 (42)	4769 (39)
Former	5394 (38)	840 (39)	4554 (38)
Current	3190 (22)	414 (19)	2776 (23)
Drinking status (n (%))			
Current	8028 (56)	888 (41)	7140 (59)
Former	3009 (21)	667 (31)	2342 (19)
Never	3211 (23)	596 (28)	2615 (22)

SD: standard deviation

Table 2

Factor loadings for general and domain-specific factor analyses: Results from ARIC NCS (N=14,252)

Cognitive test	General cognitive performance		Memory	Language	Executive function	
	Factor loading (SE)		Factor loading (SE)	Factor loading (SE)	Factor loading (SE)	
	White	Black	White+Black	White+Black	White	Black
Delayed word recall	0.49 (0.01)		0.61 (0.01)			
Logical memory	0.58 (0.01)		0.70 (0.02)			
Incidental learning	0.58 (0.01)		0.63 (0.01)			
Semantic fluency	0.66 (0.01)			0.81 (0.01)		
Boston Naming	0.57 (0.01)	0.65 (0.01)		0.65 (0.01)		
Phonemic fluency	0.50 (0.01)	0.69 (0.01)		0.69 (0.01)		
Trail Making Test, Part A	0.70 (0.01)	0.79 (0.01)			0.78 (0.01)	0.80 (0.02)
Trail Making Test, Part B	0.83 (0.01)				0.84 (0.01)	
Digit symbol substitution	0.79 (0.01)	0.86 (0.01)			0.74 (0.01)	
Digit span backwards	0.47 (0.01)	0.56 (0.02)				

Legend. Factor loadings represent correlations between an item and the underlying latent trait listed at the top of each column. Some loadings were estimated separately by race group to account for differential item functioning; see Methods.

SE: standard error

Table 3

Association between diabetes and annual change in cognitive performance: Results from ARIC NCS (N=14,252)

Cognitive domain	Factor score, adjustment for differential item functioning		Factor score, no adjustment for differential item functioning		Average of standardized scores	
	Estimate (SE)	Z	Estimate (SE)	Z	Estimate (SE)	Z
General performance	−0.064 (0.015)	−4.14	−0.058 (0.016)	−3.70	−0.041 (0.014)	−2.89
Executive functioning	−0.047 (0.016)	−2.92	−0.065 (0.015)	−4.33	−0.057 (0.013)	−4.45
Memory	--		−0.190 (0.095)	−2.01	−0.009 (0.022)	−0.39
Language	--		−0.088 (0.023)	−3.79	−0.050 (0.016)	−3.23

Coefficients represent the difference, in SD units per decade, in the rate of annual cognitive decline between persons with and without diabetes. For each general and domain-specific cognitive variable, we compared the factor score approach to an equally weighted average of standardized scores. For general cognitive performance, both the factor score and the average score used 3 tests at the 1990–92 and 1996–98 visits and 10 tests at the 2011–13 visit. The factor and average score for executive functioning both used Digit Symbol Substitution at the 1990–92 and 1996–98 visits and 3 tests (Digit Symbol Substitution, Trails A and B) at the 2011–13 visit. The factor and average score for memory both used delayed word recall at the 1990–92 and 1996–98 visits and 3 tests (delayed word recall, Logical Memory, incidental learning) at the 2011–13 visit. The factor and average score for language both used semantic fluency at the 1990–92 and 1996–98 visits and 3 tests (semantic fluency, phonemic fluency, Boston Naming) at the 2011–13 visit. Models were adjusted for age, sex, indicators for race and ARIC field center, level of education (less than high school, high school, more than high school), prevalent coronary heart disease, prevalent stroke, hypertension, total cholesterol, body mass index, cigarette history (never, former, current), apolipoprotein status (any ε4 allele or none), drinking history (never, former, current), and interactions between time and each of these variables.