Supplemental Methods

Model Building

Using the *Psych* package in R, the first model building step was an exploratory factor analysis (EFA). Eigenvalues, scree plots, a preference for parsimony, and theoretical rationale were considered when selecting the number of latent factors. Informed by the results of the EFA, confirmatory factor analysis (CFA) was used to determine the best fitting model using the "*mirt"* package in R. In the first CFA step, the seven test scores were entered as indicators without any residual covariances. Each of the seven continuous variables was recoded into an ordinal score with at least 10 observations in each response category [1, 2]. This recoding preserves the original distribution but facilitates factor score generation. Graded response models were fitted to the polytomous items using the Metropolis-Hastings Robbins-Monro (MHRM) method in mirt. Model fit indices were extracted using the M_2 function with a C_2 test of fit and quasi-monte carlo method on cases with complete data. The C_2 variant is relevant for polytomous response models lacking sufficient degrees of freedom to compute the M_2*statistic[3]. Local dependencies and fit indices, expert opinion, and knowledge of the expected covariance due to shared method variance (e.g., Trails A and B), were used to determine the best fitting, most parsimonious model. To help with model selection, our criteria for model fit were the confirmatory fit index (CFI), the Tucker Lewis Index (TLI), and the root mean squared error of approximation (RMSEA), where criteria for excellent fit were CFI> 0.95, TLI >0.95, and RMSEA<0.05 [4].

After the best fit was determined, item response theory (IRT) methods [1, 5] were used to generate factor scores for each participant in the NACC dataset by fitting a twoparameter graded response model using the *mirt* package in R. Item parameters were calibrated and saved, and scores were calculated using an expected a posteriori (EAP) method. IRT-derived scores have the important property of being invariant to the specific items used. Therefore, these scores should provide unbiased estimates of the latent trait regardless of which subtests are included; this property is particularly beneficial for retrospective research studies with missing data (as long as the missing data can be assumed to be missing at random). Latent variables were standardized with mean = 0 and SD = 1.

Shape constrained additive model (SCAM)

We first fit models allowing nonlinear corrections [6] for age and education (each constrained to be monotonic) along with an additive term for sex. Similar to the results of Kornak et al. [2], the nonlinear education effect was close to linear, so we refitted education with a linear term. Also consistent with this prior application [2], we next extracted residuals from the model, and estimated the standard deviation of the residuals at each age using a sliding window (width 11 years centered at that age). This windowing approach allowed for stable estimation in the presence of variability in the number of datapoints available at each year of age, effectively smoothing the SD estimate across an age range. A second SCAM was fitted to the estimated SD as a function of age. The resulting plot suggested that the SD was relatively constant as a function of age; we therefore fit a constant SD in the final model because these data suggest that the distribution of cognitive scores in healthy adults should not change to any clinically significant level across the age groups in this normative sample.

As a result of this model, we created a look up table which provides an estimated mean and standard deviation for each combination of age, sex, and education. This table was then used to calculate z-scores for all individuals in the validation cohort based on their score, age, sex, and education level. Because of high variability and small numbers at the youngest and oldest age ranges in the NACC dataset we used an approach to truncate extreme values when calculating factor scores for the validation sample. Specifically, for anyone below age 50, we estimate their mean and standard deviation based on that of a 50-year-old. In other words, we consider subjects at all ages below 50 to take on the model predicted score at age 50. Although the fit of the curve looks stable at ages below 50, the data are sparse. Therefore, we believe that any deviation in the fit seen below age 50 is not evidence-driven, but rather model extrapolation of the smoothness requirement in SCAM modeling. Instead of relying on model extrapolation, we used expert clinical knowledge to inform the assumption that cognitive test scores in the normative sample should not vary to any clinically significant degree between early-mid adulthood and age 50. For the same reasons, for anyone with less than 10 years of education, we estimated their scores based on a 10th grade education level. The maximum education level was set to 20 years; anyone with more than 20 years in the dataset was considered to have 20 years of education.

Other Neuropsychological Measures

Modified Trail Making Test (EF): Modified trails is a mental set-shifting task that requires subjects to serially alternate between numbers and days of the week. The task has a two-minute time limit to complete 14 correct sequences (1, Sunday, 2, Monday, 3…Saturday, 8). Completed lines per minute was used as the outcome measure for this task (e.g., 28 seconds to complete all 14 lines = 30 lines/minute).

Design Fluency (EF): Design Fluency from the Delis-Kaplan Executive Function Scale (D-KEFS) requires the participant to quickly draw designs using four straight lines that connect dots, with every design being different. Our outcome was the total correct designs on the "filled dots" condition (Trial 1) completed in 1 minute.

Letter Fluency (D-Words; EF): Participants must name as many unique words beginning with the letter "D" as quickly as they can in 1 minute. Rule violations include proper names (e.g., David, Doritos), places (e.g., Detroit), and providing the same word with different endings (e.g., drive, drives, driving). Our outcome was the number of correct, unique D-words produced in 1 minute.

Craft Story (Memory): Participants are read a short story and asked to try and repeat the story back to the examiner verbatim (maximum 44 story units). There is an immediate recall trial followed by a 20-minute delayed free recall. Our outcome was the percent retention of verbatim story units (delayed recall story units divided by immediate recall story units).

Benson Figure (Memory): Participants are asked to copy a complex geometric figure (maximum 17 points) and then, following a 10-minute delay, are asked to draw the figure from memory. Our outcome was the percent retention of figure details (delayed recall divided by copy trial).

VOSP Number Location (Spatial): Participants are shown two squares oriented vertically with the top square containing an array of numbers and the bottom square containing a single dot. Participants must indicate which number in the top square corresponds with the position of the dot in the bottom square. Our outcome was the total correct items (maximum=10).

15-Item Boston Naming Test (Language): Participants are shown line-drawing pictures of objects and asked to name the object. Pictures are arranged hierarchically by obscurity. Our outcome was the total items correct (spontaneous + semantically-cued).

Mini-Mental State Examination (MMSE)[7] : A 30-item screen of global cognition with brief assessments of orientation, attention, memory, language, and visuoconstruction. Our outcome was the total correct items (maximum=30).

Montreal Cognitive Assessment (MoCA)[8]: A 30-item screen of global cognition with brief assessments of orientation, attention, memory, language, and visuoconstruction. Our outcome was the total correct items (maximum=30).

Neuroimaging Acquisition Parameters

Magnetization prepared rapid gradient-echo (MPRAGE) sequences were used to obtained whole brain T1-weighted images (TR/TE/TI=2300/2.98/ 900 ms, α=9°; TR/TE/TI=2300/2.9/ 900 ms, α=9°). The field of view was 240x256mm, with 1x1 mm in-plane resolution and 1mm slice thickness and sagittal orientation for both sequences.

Image Processing

Before processing, all T1-weighted images were visually inspected for quality control and those with excessive motion or image artifact were excluded. Magnetic field bias was corrected using the N3 algorithm [9]. Tissue segmentation was performed using unified segmentation in SPM12 [10]. Each subject’s gray matter segmentation was warped to create a study-specific template using Diffeomorphic Anatomical Registration using Exponentiated Lie algebra (DARTEL) [11]. Subject’s native space gray and white matter segmentations were then normalized and modulated to study-specific template space using nonlinear and rigid-body transformation. Images were smoothed using a Gaussian kernel of 4-mm full width half maximum. Each subject's segmentation was carefully inspected to ensure robustness of the process.

For statistical purposes, linear and nonlinear transformations between DARTEL’s space and ICBM space were applied [12]. Quantification of volumes in specific brain regions was accomplished by transforming a standard parcellation atlas into International Consortium for Brain Mapping (ICBM) space and summing all modulated gray matter within each parcellated region of interest (ROI) [13]. Total intracranial volume was calculated for each subject as the sum of the gray matter, white matter, and cerebrospinal fluid segmentations. Putative “executive” regions included frontal grey matter, parietal grey matter, temporal grey matter, dorsolateral prefrontal cortex (DLPFC; caudal and rostral middle frontal gyrus), orbitofrontal cortex (OFC; medial and lateral orbital frontal gyrus), and anterior cingulate cortex (ACC; caudal and rostral anterior cingulate gyrus). We further assessed potential divergent validity from total occipital and pericalcarine grey matter volume.

References

1. Kramer, J.H., et al., *NIH EXAMINER: conceptualization and development of an executive function battery.* Journal of the international neuropsychological society, 2014. **20**(1): p. 11-19.
2. Kornak, J., et al., *Nonlinear Z-score modeling for improved detection of cognitive abnormality.* Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 2019. **11**(C): p. 797-808.
3. Cai, L. and S. Monroe, *A New Statistic for Evaluating Item Response Theory Models for Ordinal Data. CRESST Report 839.* National Center for Research on Evaluation, Standards, and Student Testing (CRESST), 2014.
4. Reeve, B.B., et al., *Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS).* Medical care, 2007: p. S22-S31.
5. Mungas, D., B.R. Reed, and J.H. Kramer, *Psychometrically matched measures of global cognition, memory, and executive function for assesment of cognitive decline in older persons.* Neuropsychology, 2003. **17**(3): p. 380.
6. Pya, N. and S.N. Wood, *Shape constrained additive models.* Statistics and computing, 2015. **25**(3): p. 543-559.
7. Folstein, M.F., S.E. Folstein, and P.R. McHugh, *"Mini-mental state". A practical method for grading the cognitive state of patients for the clinician.* J Psychiatr Res, 1975. **12**(3): p. 189-98.
8. Morris, J.C., *The Clinical Dementia Rating (CDR): current version and scoring rules.* Neurology, 1993. **43**(11): p. 2412-4.
9. Sled, J.G., A.P. Zijdenbos, and A.C. Evans, *A nonparametric method for automatic correction of intensity nonuniformity in MRI data.* IEEE transactions on medical imaging, 1998. **17**(1): p. 87-97.
10. Ashburner, J. and K.J. Friston, *Unified segmentation.* Neuroimage, 2005. **26**(3): p. 839-851.
11. Ashburner, J., *A fast diffeomorphic image registration algorithm.* Neuroimage, 2007. **38**(1): p. 95-113.
12. Mazziotta, J.C., et al., *A probabilistic atlas of the human brain: theory and rationale for its development.* Neuroimage, 1995. **2**(2): p. 89-101.
13. Desikan, R.S., et al., *An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest.* Neuroimage, 2006. **31**(3): p. 968-980.