NeuroImage 125 (2016) 693–704

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

The Reference Ability Neural Network Study: Life-time stability of reference-ability neural networks derived from task maps of young adults

C. Habeck [a],*, Y. Gazes [a], Q. Razlighi [a], J. Steffener [b], A. Brickman [a], D. Barulli [a], T. Salthouse [c], Y. Stern [a]

[a] Cognitive Neuroscience Division, Department of Neurology, Columbia University, NY, NY 10032, USA
[b] PERFORM Center and Department of Psychology, Concordia University, Montréal, QC H4B 1R6, Canada
[c] Department of Psychology, University of Virginia, Charlottesville, VA 22904, USA

ARTICLE INFO

Article history:
Received 12 May 2015
Accepted 26 October 2015
Available online 31 October 2015

ABSTRACT

Analyses of large test batteries administered to individuals ranging from young to old have consistently yielded a set of latent variables representing reference abilities (RAs) that capture the majority of the variance in age-related cognitive change: Episodic Memory, Fluid Reasoning, Perceptual Processing Speed, and Vocabulary. In a previous paper (Stern et al., 2014), we introduced the Reference Ability Neural Network Study, which administers 12 cognitive neuroimaging tasks (3 for each RA) to healthy adults age 20–80 in order to derive unique neural networks underlying these 4 RAs and investigate how these networks may be affected by aging.

We used a multivariate approach, linear indicator regression, to derive a unique covariance pattern or Reference Ability Neural Network (RANN) for each of the 4 RAs. The RANNs were derived from the neural task data of 64 younger adults of age 30 and below. We then prospectively applied the RANNs to fMRI data from the remaining sample of 227 adults of age 31 and above in order to classify each subject-task map into one of the 4 possible reference domains. Overall classification accuracy across subjects in the sample age 31 and above was 0.80 ± 0.18. Classification accuracy by RA domain was also good, but variable; memory: 0.72 ± 0.32; reasoning: 0.75 ± 0.35; speed: 0.79 ± 0.31; vocabulary: 0.94 ± 0.16. Classification accuracy was not associated with cross-sectional age, suggesting that these networks, and their specificity to the respective reference domain, might remain intact throughout the age range. Higher mean brain volume was correlated with increased overall classification accuracy; better overall performance on the tasks in the scanner was also associated with classification accuracy. For the RANN network scores, we observed for each RANN that a higher score was associated with a higher corresponding classification accuracy for that reference ability. Despite the absence of behavioral performance information in the derivation of these networks, we also observed some brain–behavioral correlations, notably for the fluid-reasoning network whose network score correlated with performance on the memory and fluid-reasoning tasks. While age did not influence the expression of this RANN, the slope of the association between network score and fluid-reasoning performance was negatively associated with higher ages. These results provide support for the hypothesis that a set of specific, age-invariant neural networks underlies these four RAs, and that these networks maintain their cognitive specificity and level of intensity across age.

Activation common to all 12 tasks was identified as another activation pattern resulting from a mean-contrast Partial-Least-Squares technique. This common pattern did show associations with age and some subject demographics for some of the reference domains, lending support to the overall conclusion that aspects of neural processing that are specific to any cognitive reference ability stay constant across age, while aspects that are common to all reference abilities differ across age.

Introduction

Analyses of large test batteries administered to individuals ranging from young to old, have consistently yielded latent variables, or reference abilities (RAs) that capture the majority of the variance in age-related cognitive change. Salthouse et al. have identified four domains: episodic memory, fluid reasoning, perceptual speed, and vocabulary (Salthouse, 2005, 2009; Salthouse et al., 2008). Based on these findings, Salthouse et al. have argued that a productive and efficient approach to cognitive aging research is to try to understand how aging impacts performance of this small set of RAs, rather than on specific tasks (Salthouse and Ferrer-Caja, 2003). Similarly, for cognitive neuroimaging research in aging the emphasis on age-related differences in a set of

* Corresponding author.
E-mail address: ch629@cumc.columbia.edu (C. Habeck).

http://dx.doi.org/10.1016/j.neuroimage.2015.10.077
1053-8119/© 2015 Elsevier Inc. All rights reserved.

broad neural networks underlying the reference abilities for the four cognitive domains would be more productive than a piecemeal approach focusing on separate individual tasks without consideration of commonalities between these tasks. This would allow us to more reliably explore the neural basis of aging's influence on key cognitive abilities. The Reference Ability Neural Network (RANN) Study is designed to identify networks of brain activity uniquely associated with performance across adulthood of each of the four reference abilities described above. In the RANN study, 12 tasks, three from each domain, that have reliably been associated with the corresponding RA, are administered to subjects in the scanner. Using analytic approaches that parallel those used to derive latent variables from cognitive psychometric data, we aim to determine whether four spatial fMRI networks can be derived that serve as the neural substrate for the latent cognitive structure of the reference abilities.

In a previous report (Stern et al., 2014) we introduced the RANN study and presented details of its acquisition and analysis procedures. We described an analysis intended to provide an initial representation of actual RANNs for each ability. We used a general linear model approach to summarize each subject's activation for each task into a single contrast. We then used a multivariate technique, linear indicator regression analysis, to derive four unique linear combinations of Principal Components (PC) of imaging data, one for each RA. We then investigated the ability of these constructed patterns to predict the reference domain using the activation of individual subjects for each task in held-out data. Median accuracy rates for associating component task activation with its corresponding reference ability were quite good: memory: 76%; reasoning: 82%; speed: 79%; vocabulary: 71%. We took this as an indication that it will be possible to identify unique networks associated with each reference ability.

Here we report an extension of this analysis in a larger group of participants. In our original report, we attempted to identify networks unique to each ability using data from subjects of all ages. Since the RANN study is intended to understand the sources of age-related cognitive change, it would be important to identify RANNs in younger people, and then investigate how these networks change as a function of aging. In the current study, we again used linear indicator regression analysis to derive a unique spatial covariance pattern (from a set of Principal Components) for each reference ability, but this analysis focused only on 64 individuals age 30 and below. We then investigated whether expression of these covariance patterns could successfully predict the reference domain associated with the activation of individual subjects and tasks in participants age 31 to 80. To the extent that these patterns are consistently expressed across age, this association should remain stable. However, a worsening in the ability to categorize abilities for older participants might indicate some age-related change. To the extent that we observed differences in classification accuracy, we planned to investigate the basis of these differences taking several approaches. Here we assessed whether classification accuracy 1) was lower for higher age for specific reference abilities or specific individuals, 2) was associated with the degree to which these patterns were expressed, and 3) was associated with observed age differences in mean cortical volume, cortical thickness and white-matter hyper-intensity burden. In addition to the activation particular to each reference domain, we also identified a common activation pattern in the derivation sample of participants aged 20–30. Brain-behavioral correlations and correlation with demographics was also assessed in the validation sample of participants aged 31 and above.

Material and Methods

Subjects

291 healthy adults were included in these analyses. All subjects are native English speakers, strongly right-handed, and have at least a fourth grade reading level. Subjects were screened for MRI contraindications and hearing or visual impairment that would impede testing. Subjects were free of medical or psychiatric conditions that could affect cognition. Careful screening ensured that the elder subjects did not meet criteria for dementia or Mild Cognitive Impairment (MCI). A score greater than 130 was required on the Mattis Dementia Rating Scale (Mattis, 1988). Further, performance was required to be within age-adjusted normal limits on a list-learning test, and participants were required to have no or minimal complaints on a functional impairment questionnaire (Blessed et al., 1968).

Procedure

All subjects completed screening for dementia or MCI prior to participating in the remainder of the study. They participated in two 2-hour scanning sessions. Six tasks were administered in each session in the context of fMRI studies. One session presented three Vocabulary tasks and three Perceptual Speed tasks interspersed in a fixed order: Synonyms, Digit-Symbol, Antonyms, Letter Comparison, Picture Naming, and Pattern Comparison; and the other session presented three Episodic Memory tasks and three Fluid Reasoning tasks, also interspersed in a fixed order: Logical Memory, Paper Folding, Word Order Recognition, Matrix Reasoning, Paired Associates, Letter Sets. The order of tasks within session was not varied, but the order of the two sessions was counterbalanced across subjects, with equal numbers having each order. The activation tasks were supplemented with other imaging procedures described below. At a separate session subjects completed a battery of neuropsychological tests as well as a set of questionnaires. These will not be discussed in the current report.

Stimulus presentation

Task stimuli were back-projected onto a screen located at the foot of the MRI bed using an LCD projector. Participants viewed the screen via a mirror system located in the head coil and, if needed, had vision corrected to normal using MR compatible glasses (manufactured by SafeVision, LLC. Webster Groves, MO). Responses were made on a LUMItouch response system (Photon Control Company). Task administration and collection of reaction time (RT) and accuracy data were controlled by EPrime (v2.08) running on a PC computer. Task onset was electronically synchronized with the MRI acquisition computer.

Reference Ability tasks

In the scanner, participants performed a battery of twelve computerized tasks based on the cognitive tasks that have been used to derive the RAs addressed in this report. Prior to the scan session, computerized training was administered for the six tasks included in that session. At the completion of training for each task, participants had the option of repeating the training. The tasks are described in detail in (Stern et al., 2014). For all tasks, except picture naming, responses were differential button presses. During training, responses were made on the computer keyboard and during scans they were made on the LUMItouch response system.

In the remainder of the manuscript, we will use the following shorthand notation for the reference abilities: episodic memory — MEM, fluid reasoning — FLUID, perceptual processing speed — SPEED, and vocabulary — VOCAB.

Vocabulary Tests. The primary dependent variable for all VOCAB tasks is the proportion of correct items.

Synonyms (Salthouse, 1993). Subjects have to match a given word to its synonym, or to the word most similar in meaning. The probe word is presented in all capital letters at the top of the screen, and four numbered choices are presented below.

Antonyms (Salthouse, 1993). Participants match a given word to its antonym, or to the word most different in meaning.

Picture Naming. Subjects have to verbally name pictures, adapted from the picture naming task of the WJ-R Psycho-Educational battery (Salthouse, 1998; Woodcock et al., 1989).

Perceptual Speed Tests. The primary dependent variable for all SPEED tasks is RT.

Digit Symbol. A code table is presented on the top of the screen, consisting of numbers one through nine, each paired with an associated symbol. Below the code table an individual number/symbol pair is presented. Subjects are asked to indicate whether the individual pair is the same as that in the code table using a differential button press. Subjects are instructed to respond as quickly and accurately as possible.

Letter Comparison (Salthouse and Babcock, 1991). In this task, two strings of letters, each consisting of three to five letters, are presented alongside one another. Subjects indicate whether the strings are the same or different using a differential button press.

Pattern Comparison (Salthouse and Babcock, 1991). Two figures consisting of varying numbers of lines connecting at different angles are presented alongside one another. Subjects indicate whether the figures were the same or different by a differential button press.

Fluid Reasoning Tests. The primary dependent variable for FLUID tasks is proportion of correct trials completed.

Paper Folding (Ekstrom et al., 1976). Subjects select a pattern of holes (from five options) that would result from a sequence of folds in a piece of paper, through which a hole is then punched. The sequence is given on the top of the screen, and the five options are given in a row below. Response consisted of pressing 1 of 5 buttons corresponding to the chosen solution.

Matrix Reasoning (adapted from (Raven, 1962)). Subjects are given a matrix that is divided into nine cells, in which the figure in the bottom right cell is missing. Below the matrix, they are given eight figure choices, and they are instructed to evaluate which of the figures would best complete the missing cell.

Letter Sets (Ekstrom et al., 1976). Subjects are presented with five sets of letters, where four out of the five sets have a common rule (i.e. have no vowels), with one of the sets not following this rule. Subjects are instructed to select the unique set.

Episodic Memory Tests. Note that for the MEM tasks, both the study and test phases were imaged and cannot be separated. The primary dependent variable for the memory tests is proportion of correctly answered questions.

Logical Memory. Stories are presented on the computer screen. The subject is asked to answer detailed multiple-choice questions about the story, with four possible answer choices.

Word Order Recognition. A list of twelve words is presented one at a time on the screen, and subjects are instructed to remember the order in which the words are presented. Following the word list they are given a probe word at the top of the screen, and four additional word choices below. They are instructed to choose out of the four options the word that immediately followed the word given above.

Paired Associates. Pairs of words are presented, one at a time, on the screen, and subjects are instructed to remember the pairs. Following the pairs, they were given a probe word at the top of the screen and four additional word choices below. Subjects were asked to choose the word that was originally paired with the probe word.

Image acquisition procedures

All MR images were acquired on a 3.0T Philips Achieva Magnet. There were two 2-hour MR imaging sessions to accommodate the twelve fMRI tasks as well as the additional imaging modalities. At each session, first a scout, T1-weighted image was acquired to determine patient position. All scans used a 240 mm field of view. For the EPI acquisition, the parameters were: TE/TR (ms) 20/2000; Flip Angle 72 degrees; In-plane resolution (voxels) 112x112; Slice thickness/gap (mm) 3/0; Slices 41. In addition, MPRAGE, FLAIR, DTI, ASL and a 7-minute resting BOLD scan were acquired. A neuroradiologist reviewed each subject's scans. Any significant findings were conveyed to the subject's primary care physician.

Behavioral performance variables

Behavioral performance was recorded while subjects executed the tasks in the scanner. To ensure that we included data only from instances where subjects were performing the task, we eliminated data from any task where the participant's performance was at chance or lower. For the SPEED tasks, we required accuracy of 75% or greater because the focus was on the speed of performance as represented by reaction time. Z-scores were computed for all twelve behavioral variables based on the entire study group. For the SPEED tasks, the behavioral Z-scores were reversed in sign, such that an increasing value of the behavioral Z-score implied better performance.

A small portion of scans (78 scans = 2.7% of the number of finally used scans) did not have information about behavioral performance recorded due to technical difficulties. We decided to include these scans in the analysis. The danger of type-II error, i.e. "washing out" true effects by including null observations, in our estimation, outweighed the danger of type-I error. For any brain-behavioral correlations and computations, these scans were left out.

Image analysis pre-processing procedures

Structural neuroimaging. Each subject's structural T1 scans were reconstructed using FreeSurfer v5.1 (http://surfer.nmr.mgh.harvard.edu/). The accuracy of FreeSurfer's subcortical segmentation and cortical parcellation (Fischl et al., 2002, 2004) has been reported to be comparable to manual labeling. Each subject's white and gray matter boundaries as well as gray matter and cerebral spinal fluid boundaries were visually inspected slice by slice, manual control points were added in the case of any visible discrepancy, and reconstruction was repeated until we reached satisfactory results within every subject. The subcortical structure borders were plotted by FreeView visualization tools and compared against the actual brain regions. In case of discrepancy, they were corrected manually. Finally, we computed mean values for 68 cortical regions of interests (ROIs) for cortical thickness and cortical volume for each participant to be used in group-level analyses.

White-matter hyper-intensities were obtained from FLAIR images according to the protocol outlined by Brickman et al. (2011).

Functional neuroimaging. Each individual's 12 fMRI scans were preprocessed in the same manner using the FSL software package (Smith et al., 2004). The processing of the functional imaging data involved the following basic steps: 1) within-subject histogram computation for each subject volume to identify noise (FEAT); 2) subject-motion correction (MCFLIRT); 3) slice-timing correction; 4) brain-mask creation from first volume in subject's fMRI data; 5) high-pass filtering (T = 128 sec); 6) pre-whitening; 7) General-Linear-Model (GLM) estimation with equally temporally filtered regressors and double-gamma hemodynamic response functions; 8) registration of functional and structural images with subsequent normalization into MNI space (FNIRT).

GLM for each subject and each task consisted of block-based time-series analysis for SPEED, MEM, and VOCAB tasks and event-related modeling for FLUID tasks (to separate out correct and incorrect trials) using FEAT in FSL. For group analysis, contrary to the usual voxel-wise FSL practice, we obtained standardized contrast images for every subject and task to perform group-level multivariate analysis (next section). Contrast images captured all brain activation pertinent to all cognitive processes present in the task in a broad contrast of "task performance vs fixation cross"; there was no separation of stimulus presentation and behavioral response in our task design, which would have been prohibitive in terms of complexity and time.

Derivation of RANNs with linear-indicator regression in participants up to age 30

This analysis intended to use a multivariate approach to derive 4 RANNs that were best associated with the 3 tasks in each RA. We used a linear-indicator regression approach (Hastie et al., 2009). This approach decomposes activation in each task to a set of PCs and then derives the optimal combination of PCs that discriminates between the 3 tasks in a RA and the other 9 tasks. By design, this analysis was restricted to 64 participants up to 30 years in age. These 64 participants accounted for 593 subject-and-task parametric maps.

First, a Principal Components Analysis was run on the 593 maps, and the individual pattern scores, or Subject Scaling Factors (SSF), for the first 200 PCs were obtained by an inner product of all 200 PCs with the 593 maps. Concretely, the pattern score matrix SSF is computed with the following multiplication,

$$\mathbf{SSF}\ (\mathrm{i},\mathrm{k}) = \mathbf{Y}(:,\mathbf{i})^{\mathrm{t}}\,\mathbf{V}(:,\mathrm{k})$$

where *i* denotes the subject-task index and runs from 1 to 593, Y(:,i) represents one activation map, i.e. the *i*th column in matrix **Y**, and *k* indicates the PC index, running from 1 to 200. V(:,k) is this the *k*th column in the matrix of Principal Components, **V**. Both matrices **Y** and **V** have as many rows as voxels in the brain. Selected columns of the array **SSF** were then used as independent variables in a subsequent linear-indicator regression (Hastie et al., 2009) to predict an indicator matrix **I**. **I** had 593 rows and 4 columns, and places a value of 1 in the appropriate column depending on the RA domain that the task-map belongs to and was indexed by the row position. Summing over all entries in **I** recovers the total number of maps in the analysis: 593. The regression equation can be written as

$$\mathbf{I} = [\mathbf{SSF}(:,1:\mathrm{k})\ \ \mathbf{1}]\ \ \mathbf{B}\ \ + \mathrm{error}$$

where **SSF**(:,1:k) is the array of pattern scores for the first *k* PCs, and **1** denotes an intercept term. **B** is an array of regression weights of format $(k+1) \times 4$. The corresponding 4 RANNs were constructed by applying the regression weights to the PCs according to

$$\mathbf{RANN} = \mathbf{V}(:,1:\mathrm{k})\ \ \mathbf{B}(:,1:4)$$

To select an optimal set of PCs, i.e. the best number *k* of included PCs, we used a goodness-of-fit measure, the AIC criterion (Burnham and Anderson, 2002), computed for each of the 4 dependent variables in the indicator matrix to arrive at an average value for each set of PCs. AIC picks an optimal bias-variance tradeoff and minimizes the residual sum of squares, while keeping the number of parameters in the model at a minimum. We picked *k* according to the AIC criterion, i.e. we varied *k* from 1 to 200, running the linear-indicator regression each time, and chose *k* such that AIC was minimal. For the case that several very similar local minima in the AIC curve were present, we decided beforehand to take the set with the minimum number of PCs, to keep the variance contribution in the data as large as possible.

Once k was determined, we performed the linear-indicator regression for the full sample and computed the RANNs. To determine the robustness of RANN voxel-loadings, we performed a semi-parametric bootstrap resampling procedure (Efron and Tibshirani, 1998) with 500 iterations, which resampled from the full 593 scans with replacement, each time performing the derivation of the RANNs. The variability of the voxel loadings in the bootstrap resampling procedure around the point estimate values can be approximated as a Z value at voxel location *i* according to the formula

$$\mathrm{Z(i)} = \mathrm{RANN(i)}\ /\ \mathrm{bootstrap{-}STD(i)}$$

Robust loadings fulfill $|Z| > 3$ and are visualized in the four RANN images.

Investigating the ability of RANNS to Predict RA domains in participants > age 30

Next, we investigated the ability of the RANNs derived in this younger group to predict the underlying RA domain for any individual subject's activation from the subjects aged 31–80.

For any scan **y**, a prediction of the reference label can be made according to

$$\mathrm{L} = [\mathbf{y}\,\mathbf{V}(:,1:\mathrm{k})'\ \ \mathbf{1}]\ \ \mathbf{B}.$$

L is 1 x 4 row vector and contains the degree to which the scan loads onto each RANN, while V and B have already been determined from the younger subjects' data. The biggest loading determines the predicted reference-domain label. The metric chosen for quantifying classification performance was mean prediction accuracy, computed as the proportion of hits for each reference ability. Overall classification accuracy, as well as classification accuracy for each ability was calculated.

We then explored potential correlates of classification accuracy by correlating it with structural, performance and demographic covariates.

Computing subject expression of the RANNs

For every one of the 291 participants, the 4 expression scores were calculated by computing the inner product of each RANN with its corresponding task maps (up to 3) for that participant, and averaging the expression values across tasks. For instance, if we assume that a participant has all 3 task maps for the MEM domain available (i.e. Logical Memory, Word Order and Pairs Associates) and these 3 maps are assembled in a matrix **Y** that has as many rows as voxels, and 3 columns, and the MEM–RANN is represented by a column vector, **v**, we can compute a 3-row expression vector according to the inner product

$$\mathbf{Y'v}$$

with a subsequent average across all 3 tasks, to arrive at a single score. This is done in analogous fashion for the other reference domains as well. If only 2 of the 3 tasks of a reference domain are present, the average is performed across the subset of 2 tasks. If only one task is available, no averaging is necessary. If no task is present for this reference domain in this participant, no score can be computed and the participant is left out of any brain-behavioral correlations.

Derivation of common task-activation pattern in participants up to age 30

Since many cognitive processes related to stimulus presentation and behavioral response are likely to be common to all 12 tasks with substantial variance contributions and possible age effects, we decided to derive a common activation pattern as well. The derivation sample from which this common activation pattern was derived was the same as in Section 2.2.6, i.e. all 593 task-activation maps of participants between the ages of 20 and 30. A simple mean-contrast Partial Least Squares (PLS) routine (McIntosh et al., 1996; McIntosh and Lobaugh, 2004) was employed, i.e. 12 mean-contrast maps were computed, one per task, and then submitted to a PCA. The first Principal Component was taken as the point estimate of the common activation pattern. Robustness of voxel loadings was again assessed with a bootstrap procedure (Efron and Tibshirani, 1998) with 500 iterations.

Pattern scores of the common pattern are computed in a manner identical to the approach outlined in Section 2.2.8, and result in one score per participant per reference domain. Brain–behavioral correlations are likewise performed in an identical manner to 2.2.8.

Beyond these brain–behavioral correlations, we can also obtain pattern scores separately for each task, and ask whether tasks that belong to the same reference domain show a higher common-pattern-score correlation than tasks belong to different reference domains. We have 4 x 3 = 12 pairings of tasks within reference domain, and 6 x 9 = 54 pairings between reference domains. The average difference in Fisher-Z correlation can be used as a statistic to ascertain convergent and discriminant validity, i.e. tasks belonging to the same reference domain should display high correlation, while tasks belonging to different domains should display low correlation, making the overall difference

$$\Delta Z = \quad Z(\text{within}) – Z(\text{between})$$

as large as possible. ΔZ can be used as a statistic in a permutation test — (1) to ascertain construct validity per se, and (2) to compare the two age groups, and check whether the older participants show a lower value of ΔZ, interpretable as 'dedifferentiation' in the deployment of the common activation pattern. For (1) generation of the null distribution implies a permutation of reference-domain labels of the 12 tasks; for (2) the null distribution is generated by permuting participants between the derivation and validation sample.

Age-specific derivation of common pattern

In addition to ascertaining how the common pattern derived in young people manifests in middle-aged and older adults, we can perform the pattern derivation for each age decade separately via the mean-contrast PLS analysis steps outlined in the previous section.

For each pattern we can read off the percentage variance accounted for (%VAF) simply by the magnitude of the first Eigen value, and check whether %VAF shows any association with age. Further, we can also compute the spatial correlation coefficients (Fisher-Z) between the common patterns of different age decades. The maximal age gap is 5 decades, with one correlation coefficient between decades 2 and 7. For an age gap of 4 decades, we have 2 correlation coefficients: the correlation between decades 2 and 6, and between decades 3 and 7. We can average all available correlation coefficients for all 5 age gaps and plot the strength of the correlation against the age gap.

For the relationships between %VAF and age, as well as spatial correlation and age gap, we can perform inferential tests of linear trends with permutation tests of 1,000 iterations, for which participants will be randomly shuffled between age decades. For the resulting %VAF and spatial correlation of the ensuing null patterns we will compute the linear trend again to generate null distributions. Two-tailed tests will check whether the point estimate linear trends in both %VAF and spatial correlation fall in the tail of the null distributions, and approximate the p-level as the fraction of iterations causing more extreme slope parameters as in the point estimate.

Results

Subject Demographics

Demographic features of the study participants are summarized in Table 1.

Scanning statistics are shown in Fig. 1. We allowed incomplete data sets and utilized every admissible scan in our analysis. For the 291 participants, there were 3,137 parametric task maps. Applying stringent screens and demanding above-chance performance in the scanner eliminated 193 maps (~6%), leaving 2,944 task-subject maps for analysis.

Linear-indicator regression analysis

RANN Derivation

Linear-indicator regression analysis was used to derive 4 separate spatial covariance patterns associated specifically with each of the four RAs in participants of age 30 or below. The first 91 PCs were chosen to construct the four RANN patterns because they yielded the global minimum of the Akaike Information Criterion. The four RANNs are illustrated in Fig. 2 and described in Tables 2–5.

Classification accuracy in participants in age ranges 20–30 and 31–80

Since our RANNs were derived in the subsample of participants of age 30 and younger, we separated this derivation sample from the validation sample of ages 31–80. We computed overall classification accuracy, and classification accuracy broken down by reference domain for derivation and validation samples separately. In the derivation sample the overall classification accuracy was 0.93, while being lower, as expected, in the validation sample at 0.81. The relationship between actual vs predicted RA is presented in Table 6. The accuracy for specific abilities in the derivation sample was as follows: MEM hit rate was 0.88; FLUID hit rate was 0.93; SPEED hit rate 0.91; VOCAB hit rate 0.98. Again, the accuracies for specific abilities in the validation sample were lower, but still very good: MEM hit rate was 0.72; FLUID hit rate was 0.78; SPEED hit rate 0.79; VOCAB hit rate 0.94. The full confusion matrices of the RANN application in both derivation and validation samples are shown in Table 6.

We also calculated classification accuracy for each individual participant in both derivation and validation samples in order to examine the distribution of accuracy across the age range and compile descriptive statistics across subjects. Overall classification accuracy across subjects in the validation samples (age 31–80) was 0.80 ± 0.18; broken down by reference domain, we have: MEM = 0.72 ± 0.32; FLUID = 0.75 ± 0.35; SPEED = 0.79 ± 0.31; VOCAB = 0.94 ± 0.17. Classification accuracy by decade is illustrated in Fig. 3. There is no trend for reduced classification accuracy with higher ages. In addition, a one-sample T-test for the difference from chance performance (=0.25 accuracy) was highly significant for all decades (p ~ e − 17), indicating that classification accuracy remained good in each decade.

The classification accuracy in the derivation sample (age 20–30) was, as expected, substantially higher: overall accuracy = 0.92 ± 0.11; MEM = 0.88 ± 0.25; FLUID = 0.91 ± 0.24; SPEED = 0.92 ± 0.17; VOCAB = 0.98 ± 0.14 (no figure shown).

Correlates of classification accuracy

We next assessed correlates of classification accuracy only for the participants in the validation sample of age 31 or greater. Both overall classification accuracy and classification accuracy for individual reference abilities were considered. Table 7 summarizes the correlation coefficients and p-values of bivariate relationship between classification accuracy and a variety of demographic, cognitive, and neural measures.

There was no significant relationship with age for any of the classification accuracy measures. Education was not associated with better classification accuracy. Higher NART IQ was associated with better overall classification and for MEM. Higher DRS score was associated with

Table 1
Participant demographics and brain measures.

	Age 20–29	Age 30–39	Age 40–49	Age 50–59	Age 60–69	Age 70–79
N	60	53	41	49	45	43
Sex	20 M, 40 F	19 M, 34 F	23 M, 18 F	25 M, 24 F	24 M, 21 F	21 M, 22 F
Education (years)	15.7 ± 2.1	16.3 ± 2.6	15.9 ± 2.6	15.8 ± 2.1	16.2 ± 2.6	17.6 ± 2.5
DRS total	140.1 ± 2.5	139.8 ± 2.6	139.4 ± 2.8	140.4 ± 3.1	139.7 ± 2.9	139.3 ± 2.9
AmNART IQ	113.0 ± 7.7	110.9 ± 9.0	115.4 ± 8.3	115.2 ± 8.9	117.5 ± 9.9	121.4 ± 6.5
WMH	1.47 ± 1.46	1.40 ± 1.43	1.02 ± 0.73	1.47 ± 1.49	3.21 ± 4.13	3.37 ± 3.17
Mean cortical ROI volume	7,520 ± 687	7,105 ± 621	7,183 ± 716	6,903 ± 663	6,507 ± 575	6,465 ± 570
Mean cortical ROI thickness	2.70 ± 0.10	2.65 ± 0.09	2.65 ± 0.09	2.59 ± 0.09	2.51 ± 0.10	2.49 ± 0.11

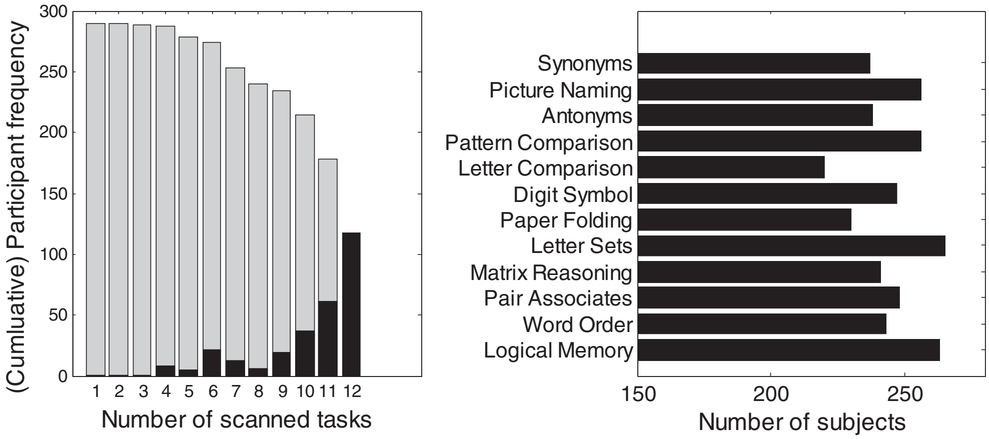

Fig. 1. Left panel: histogram and cumulative distribution of subject numbers displayed by the number of tasks that were completed. Only 117 subjects had all 12 tasks completed as can been seen from the frequency distribution (black color), thus insistence on completeness would have cut down on the available data substantially. 215 subjects had at least 10 tasks completed as can be seen from the cumulative distribution (gray color). Right panel: number of subjects for each of the 12 tasks. The least populated task was 'Letter Comparison' with 220 subjects, the most populated task was 'Letter Sets'. We checked whether missingness was associated with age, years of education, gender or DRS score, but found no significant relationship.

better overall classification accuracy and for MEM and SPEED. Mean cortical thickness was not significantly related to classification accuracy. Higher mean brain volume was correlated with better overall classification accuracy, and for MEM and FLUID. Higher expression of all RANN networks but VOCAB was associated with better overall classification accuracy. Higher expression of any particular RANN was strongly associated with the corresponding classification accuracy for that reference ability. The SPEED and VOCAB-RANNs though, seemed to interfere with each other's domain classification, and RANN pattern scores were negatively associated with each other's classification accuracy.

More interestingly, classification accuracies were further positively correlated with participants' behavioral task performance in the scanner. Overall classification accuracy was associated with overall performance, and with performance in each task except for the SPEED domain. FLUID classification accuracy was related to overall performance, MEM and FLUID performance. SPEED classification accuracy was related to overall performance, as well as MEM and VOCAB performance.

Correlates of RANN Expression

We also assessed correlates of RANN expression, quantified with RANN pattern scores, across all participants. The correlation and p-values for these bivariate correlations are summarized in Table 8. There was no correlation of the expression of any RANN pattern score with age, years of education, cortical thickness or white-matter hyperintensity burden. Mean gray matter volume was correlated with expression of the MEM- and FLUID-RANN in the expected direction: higher mean volume was associated with higher expression. Verbal

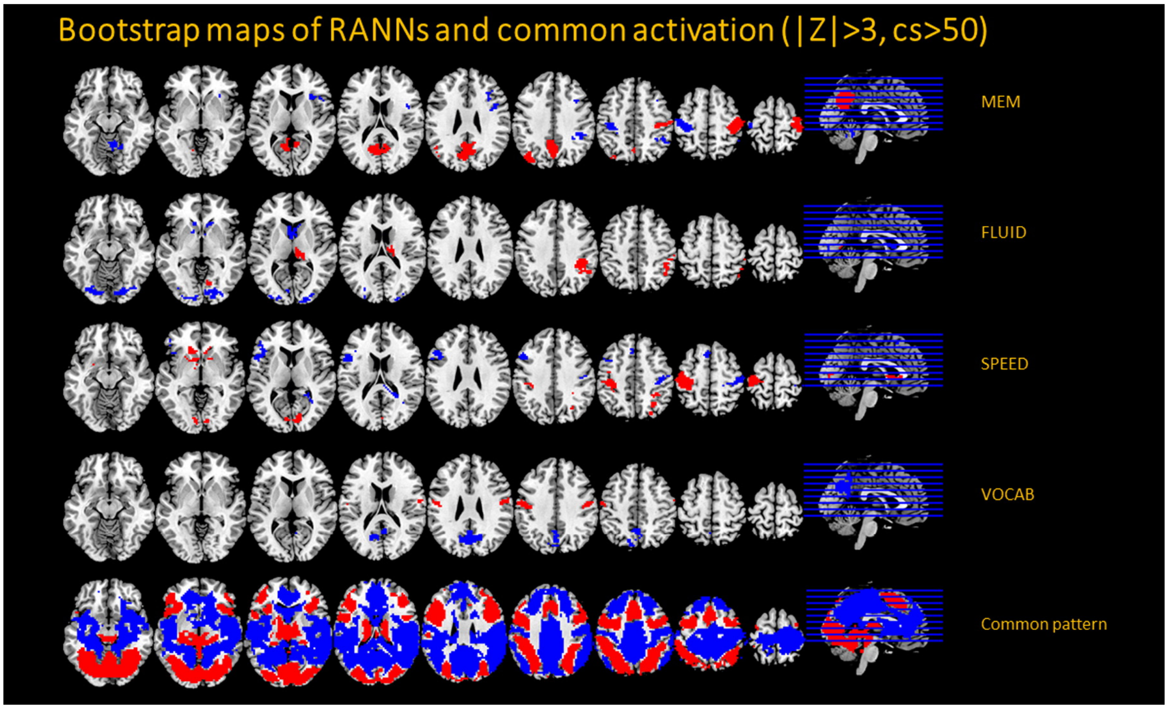
Bootstrap maps of RANNs and common activation (|Z|>3, cs>50)
MEM
FLUID
SPEED
VOCAB
Common pattern

Fig. 2. Illustration of areas with robust voxel loadings for all RANNs for each ability and for the common activation pattern obtained from mean-contrast PLS. Areas with covarying increases in activation are represented in red, while those with covarying decreases in activation are represented in blue. Only areas surviving bootstrap procedures are presented.

Table 2
Areas of activation (positive loadings) and de-activation (negative loadings) for MEM-RANN, obtained with bootstrap resampling procedure. |Z| > 3, cluster size (CS) >50.

X	Y	Z	CS	Z	AAL label
Positive Loadings					
−6	−69	24	645	8.7408	Calcarine_L
−3	−54	36	645	5.7274	Precuneus_L
−9	−60	12	645	5.4053	Calcarine_L
6	−63	24	645	5.1229	Precuneus_R
9	−54	12	645	4.9055	Calcarine_R
−6	−48	6	645	4.859	Calcarine_L
−9	−66	0	645	3.1422	Lingual_L
42	−24	57	360	6.0478	Postcentral_R
30	−24	75	360	5.2408	Precentral_R
57	−18	54	360	4.6555	Postcentral_R
27	−12	66	360	4.1101	Precentral_R
−33	−81	39	130	5.2768	Occipital_Mid_L
−45	−57	33	130	4.0188	Angular_L
−39	−69	42	130	4.0106	Angular_L
−48	−60	21	130	3.2431	Temporal_Mid_L
Negative Loadings					
−36	−27	51	227	−6.4731	Postcentral_L
−45	−18	54	227	−5.9614	Postcentral_L
−36	−24	66	227	−4.0501	Precentral_L
−18	−30	54	227	−3.6212	No AAL label
15	−54	−18	225	−6.2719	Cerebellum_4_5_R
6	−63	−21	225	−5.7648	Vermis_6
21	−45	−24	225	−5.047	Cerebellum_4_5_R
0	−51	−21	225	−3.8268	Vermis_4_5
30	−42	42	73	−4.5135	No AAL label
48	−48	54	73	−4.176	Parietal_Inf_R
45	−42	42	73	−3.7673	SupraMarginal_R
33	21	9	56	−5.6581	Insula_R
51	18	6	56	−3.3929	Frontal_Inf_Oper_R
30	24	−3	56	−3.371	Insula_R
33	24	27	54	−4.5347	Frontal_Inf_Tri_R
36	15	42	54	−3.8232	Frontal_Mid_R
30	12	30	54	−3.1608	Frontal_Inf_Oper_R
45	3	24	51	−3.844	Frontal_Inf_Oper_R
42	18	21	51	−3.4584	No AAL label

Table 3
Areas of activation (positive loadings) and de-activation (negative loadings) for FLUID-RANN, obtained with bootstrap resampling procedure. |Z| > 3, cluster size (CS) >50.

−X	Y	Z	CS	Z	AAL label
Positive Loadings					
36	−36	39	189	4.8296	No AAL label
45	−48	54	189	4.2708	Parietal_Inf_R
51	−33	48	189	3.6827	SupraMarginal_R
54	−51	39	189	3.4209	Parietal_Inf_R
6	−18	12	89	4.3411	Thalamus_R
18	−27	12	89	4.0694	Thalamus_R
15	−12	18	89	3.2021	Thalamus_R
12	−69	−3	51	4.1808	Lingual_R
Negative Loadings					
15	−87	−9	224	−7.3552	Lingual_R
18	−96	6	224	−6.4772	Occipital_Sup_R
39	−78	−12	224	−5.0322	Occipital_Inf_R
27	−93	18	224	−4.5458	Occipital_Mid_R
33	−90	−6	224	−3.7444	Occipital_Inf_R
−21	−93	9	186	−5.2011	Occipital_Mid_L
−18	−87	−9	186	−5.0613	Lingual_L
−12	−96	−3	186	−4.9686	Calcarine_L
−33	−90	6	186	−4.4834	Occipital_Mid_L
−33	−78	−15	186	−4.4726	Fusiform_L
−6	−84	6	186	−3.7696	Calcarine_L
−27	−78	21	186	−3.5347	Occipital_Mid_L
−27	−93	21	186	−3.4097	Occipital_Mid_L
−3	9	9	151	−5.0304	No AAL label
9	18	0	151	−3.9943	Caudate_R
−12	24	0	151	−3.8638	Caudate_L
6	6	0	151	−3.7273	Caudate_R

Table 4
Areas of activation (positive loadings) and de-activation (negative loadings) for SPEED-RANN, obtained with bootstrap resampling procedure. |Z| > 3, cluster size (CS) >50.

X	Y	Z	CS	Z	AAL label
Positive Loadings					
−42	−21	54	390	7.244	Postcentral_L
−42	−9	48	390	3.792	Postcentral_L
−36	−39	51	390	3.399	Parietal_Inf_L
−15	27	−6	172	4.615	Caudate_L
−6	6	0	172	4.556	No AAL label
−24	15	−6	172	3.919	Putamen_L
−3	18	0	172	3.872	Caudate_L
12	24	−6	172	3.864	Caudate_R
18	12	−6	172	3.643	Putamen_R
12	−81	0	143	6.061	Lingual_R
−9	−87	0	143	6.015	Calcarine_L
3	−81	15	143	3.778	Calcarine_L
30	−60	51	79	4.711	Parietal_Sup_R
36	−45	48	79	4.207	Parietal_Inf_R
24	−69	45	79	3.890	Occipital_Sup_R
Negative Loadings					
−48	15	36	324	−5.005	Frontal_Inf_Oper_L
−48	33	12	324	−4.908	Frontal_Inf_Tri_L
−54	15	6	324	−4.631	Frontal_Inf_Oper_L
−54	12	18	324	−4.219	Frontal_Inf_Oper_L
−42	21	−3	324	−3.758	Frontal_Inf_Orb_L
42	−24	54	145	−6.511	Postcentral_R
24	−30	60	145	−4.568	Postcentral_R
45	−30	69	145	−3.448	Postcentral_R
33	−30	69	145	−3.342	Postcentral_R
42	−15	33	145	−3.309	Postcentral_R
21	−42	15	70	−4.379	No AAL label
30	−54	12	70	−4.213	Calcarine_R
−3	21	54	51	−4.987	Supp_Motor_Area_L
−3	24	42	51	−3.530	Frontal_Sup_Medial_L

intelligence (NARTIQ) was *negatively* correlated with the VOCAB-RANN score.

With regard to the relationship between a RANN and performance of the respective reference ability, only FLUID correlated significantly with expression of the FLUID-RANN; the correlation between network expression score and performance for MEM was marginal. There was a negative correlation between RANN expression and performance in the VOCAB domain, but on closer inspection this negative correlation was found to be caused by one influential data point, and thus cannot be considered robust. Several cross-domain correlations were noted: expression of the SPEED-RANN was positively correlated with MEM and VOCAB performance, and FLUID-RANN was also positively correlated with MEM.

While age did not influence the expression of the FLUID-RANN in participants age 31 and above, the natural question arises whether the brain-behavioral relationship between the FLUID-RANN expression and performance is moderated by age. Using the FLUID-RANN score,

Table 5
Areas of activation (positive loadings) and de-activation (negative loadings) for VOCAB-RANN, obtained with bootstrap resampling procedure. |Z| > 3, cluster size (CS) >50.

X	Y	Z	CS	Z	AAL label
Positive Loadings					
−51	−9	33	138	5.0676	Postcentral_L
−42	−18	36	138	4.7497	Postcentral_L
48	−12	36	96	5.388	Postcentral_R
57	−6	45	96	3.5301	Precentral_R
Negative Loadings					
0	−72	33	410	−5.1818	Cuneus_L
−15	−66	21	410	−5.0642	Cuneus_L
9	−63	24	410	−4.3016	Precuneus_R
0	−54	51	410	−4.0942	Precuneus_L
6	−54	12	410	−3.9938	Calcarine_R
−12	−72	45	410	−3.7944	Parietal_Sup_L

Table 6
Across-subjects confusion matrix of RA label prediction in participants age 31–80 (validation sample) and in participants age 20–30 (derivation sample). The overall accuracy of classification was 0.81 in the age range 31–80, and 0.93 in the age range 20–30.

Predicted (age 31–80)				
Actual	**MEM**	**FLUID**	**SPEED**	**VOCAB**
MEM	0.72	0.07	0.03	0.18
FLUID	0.08	0.78	0.05	0.09
SPEED	0.01	0.02	0.79	0.18
VOCAB	0.02	0	0.04	0.94
Predicted (age 20–30)				
MEM	0.88	0.03	0.03	0.06
FLUID	0.01	0.93	0.04	0.02
SPEED	0	0	0.91	0.09
VOCAB	0.01	0	0.01	0.98

age and their interaction to predict FLUID performance indeed yields a significant interaction at p = 0.0059. We also broke participants into age decades (3–7) and plotted 10,000 bootstrap simulations of the brain-behavioral correlation for each decade (Fig. 4). Decreasing linear age trends were computed for each sample, the one-tailed p-level of these age trends was revealed to be p = 0.0064. The relationship to mean gray-matter or MEM performance did not reveal any formal interaction with age (results not shown).

Common activation pattern

A common activation pattern was derived in the derivation sample (ages 20–30). Topographic loadings are available in Fig. 2, and Supplementary Table S1. As expected, areas of activation were largely congruent with previous reports of the dorsal-attentional or task-positive network, while the negative loadings were strongly reminiscent of the default-mode network (Elton and Gao, 2015; Raichle, 2015).

In contrast to the RANNs, expression of the common activation pattern in the different reference domains showed associations with age: the pattern score was positively correlated with age for MEM and SPEED (and marginally with VOCAB), while being negatively correlated with age for the FLUID domain.

The common-pattern score in the SPEED domain also correlated positively with years of education and VOCAB performance, while correlating negatively with the DRS score. For none of the reference domains did the common pattern score correlate with the respective performance variable.

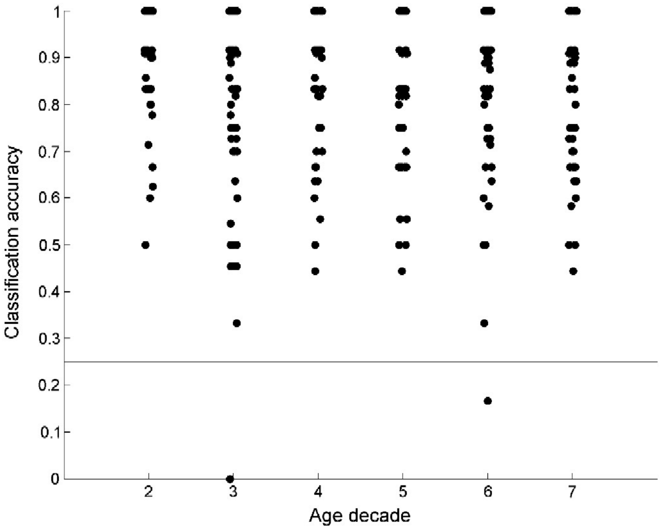

Fig. 3. Classification accuracy suggests very good generalization of RANNs beyond the derivation sample. A one-sample T-test for the difference from random performance (indicated by line at accuracy = 0.25) is highly significant for all decades (3–7) at p = 10e−21.

In addition to these somewhat difficult-to-reconcile findings, we can ask whether good convergent and discriminant validity is manifested by the pattern score, as outlined in Section 2.2.9: does usage of the common pattern appear more similar in tasks chosen from the same reference domain? This appears to be the case: the validity statistic ΔZ which captures the difference in pattern-score correlation for tasks of the same vs. different reference domains yields point estimates of ΔZ = 0.27 and ΔZ = 0.32, for the derivation and validation samples, respectively. Permutation tests were performed to assess statistical significance in each sample. Both samples' ΔZ values were highly significant at $p < 0.0001$. For an age-related comparison between samples, we performed another permutation test, this time permuting participants between the samples, rather than randomizing the reference-label assignments. This comparison did not reveal any difference between the samples (p = 0.77), suggesting that the cognitive specificity of the common-pattern deployment is unchanged between age groups.

Closer visual inspection of the 12 x 12 correlation matrices for both samples in Fig. 5 shows generally good convergent and discriminant validity for all reference domains, with some exceptions: (1) discriminant validity between SPEED and VOCAB domains is poor in both derivation and validation samples as some task pairings of SPEED and VOCAB tasks yield unduly high correlations; (2) convergent validity for the MEM and VOCAB domains in the derivation sample of younger participants is somewhat reduced as Word Order and Picture Naming tasks do not correlate highly with the other two tasks in their respective domains, thereby breaking domain convergence. As the formal age comparison confirms, though, there are no appreciable differences in overall construct validity between the derivation and validation sample.

Lastly, we re-derived the common activation pattern for each decade separately, tracking the variance accounted for by the pattern as well as the spatial similarity between patterns as a function of age gap.

Fig. 6 reveals how %VAF behaves a function of age decade. A rough linear age trend with an undershoot and overshoot by decade 3 and 4, respectively, is visible. The spatial correlation between patterns of different age decades shows a linear decline with the age gap, as expected.

Permutation tests of 1,000 iterations gave two-tailed p-levels that are statistically significant and confirmed the increasing linear trend of %VAF with age decade (p = 0.0130), and the decreasing linear trend of spatial-pattern similarity with the age gap in decades ($p < 0.001$).

Discussion

The Reference Ability Neural Network (RANN) Study is designed to identify networks of brain activity uniquely associated with performance of each of the four reference abilities across adulthood, and then to explore potential influences on these RANNs that might in turn help explain age-related changes in performance. This paper contributes to this goal by exploring how the neural substrates underlying each of these abilities is distinct, and how biological aspects of aging may influence the integrity and distinctiveness of these processes.

In a previous study, on a smaller subset of the participants considered here, we used indicator regression analysis to derived four unique activation patterns, one for each of the four reference abilities. In that analysis, 174 subjects of all ages were included. In held out data we were able to show that given the relative expression of the four activation patterns in individuals' fMRI data from one of the 12 tasks, we could quite accurately identify the underlying reference ability. Here we extended this analysis by reasoning that the specific network underlying every reference ability should be most intact in individuals at young age. Thus, this is the optimal age range in which to identify these networks. Further, this would allow us to determine whether these networks remain intact across age. There is reason to believe that this might not be the case: various aspects of age-related brain changes could likely damage the integrity of the brain networks. Further, the de-differentiation hypothesis would predict that specificity of the

Table 7
Relationship of several variables to the classification accuracy of reference-domain labels in participants age 31 and above, based on the networks derived in participants age 20–30. Both overall classification accuracy and classification accuracy for individual reference domains are considered. Correlation coefficients and p-values are listed. Bolded cells are statistically significant at p ≤ 0.05.

Classification Accuracy for RA label						
Predictors	Overall	MEM	FLUID	SPEED	VOCAB	
Age	R = 0.03; p = 0.62	R = 0.02; p = 0.73	R = −0.10; p = 0.15	R = −0.06; p = 0.41	R = 0.02; p = 0.80	Subject demographics
Education	R = 0.12; p = 0.06	R = 0.07; p = 0.30	R = 0.11; p = 0.09	R = 0.12; p = 0.07	R = −0.08; p = 0.20	Structural brain measures
NARTIQ	**R = 0.21; p = 0.002**	R = 0.04; p = 0.52	R = 0.10; p = 0.16	**R = 0.24; p = 0.001**	R = 0.01; p = 0.84	
DRS Score	**R = 0.18; p = 0.007**	R = 0.13; p = 0.07	**R = 0.15; p = 0.03**	**R = 0.21; p = 0.003**	R = −0.05; p = 0.47	
Mean Volume	**R = 0.19; p = 0.006**	**R = 0.15; p = 0.04**	**R = 0.26; p = 0.0003**	R = 0.08; p = 0.28	R = 0.0003; p = 0.99	
Mean Thickness	R = 0.01; p = 0.87	R = −0.04; p = 0.53	R = 0.13; p = 0.06	R = 0.005; p = 0.94	R = −0.07; p = 0.33	
WMH	R = 0.10; P = 0.17	R = 0.06; p = 0.45	R = 0.03; p = 0.70	R = 0.08; p = 0.26	R = 0.01; p = 0.84	
Common pattern	R = 0.09; p = 0.41	R = −0.10; p = 0.33	R = 0.10; p = 0.36	R = 0.13; p = 0.21	R = 0.10; p = 0.33	fMRI Pattern scores
MEM-RANN	**R = 0.44; p < 0.0001**	**R = 0.68; p < 0.0001**	R = 0.02; p = 0.70	**R = 0.19; p = 0.01**	R = −0.06; p = 0.47	
FLUID-RANN	**R = 0.33; p < 0.0001**	R = 0.04; p = 0.53	**R = 0.58; p < 0.001**	R = 0.07; p = 0.32	R = −0.06; p = 0.41	
SPEED-RANN	**R = 0.32; p < 0.0001**	R = 0.02; p = 0.71	R = 0.12; p = 0.10	**R = 0.56; p < 0.0001**	**R = −0.20; p = 0.003**	
VOCAB-RANN	R = −0.02; p = 0.77	R = −0.04; p = 0.43	R = 0.09; p = 0.19	**R = −0.30; p < 0.0001**	**R = 0.49; p < 0.0001**	
MEM-Perf	**R = 0.23; p = 0.0007**	R = 0.11; p = 0.12	**R = 0.26; p = 0.0001**	**R = 0.19; p = 0.008;**	R = −0.02; p = 0.70	Behavioral performance scores
FLUID-Perf	**R = 0.27; P < 0.0001**	R = 0.11; p = 0.11	**R = 0.31; p < 0.0001**	**R = 0.18; p = 0.01**	R = −0.04; p = 0.56	
SPEED-Perf	R = 0.12; p = 0.08	R = −0.15; p = 0.03	R = 0.09; p = 0.24	R = −0.13; p = 0.06	R = 0.02; p = 0.77	
VOCAB-Perf	**R = 0.20; p = 0.002**	R = 0.02; p = 0.70	R = 0.12; p = 0.09	**R = 0.21; p = 0.002**	R = 0.09; p = 0.17	
Overall Perf	**R = 0.35; p < 0.0001**	R = 0.13; p = 0.06	**R = 0.30; p < 0.0001**	**R = 0.28; p < 0.0001**	R = 0.05; p = 0.17	

estimation underlying any particular reference ability might be reduced with aging.

Therefore in the current study we again conducted a linear-indicator regression analysis but limited it to the 64 study participants age 20 to 30. Within this group we were able to derive patterns that were uniquely associated with each of the four reference abilities. We were now able to apply these candidate RANNs to data from adults age 31 and above. It is important to note that the data from the older subjects had no part in the derivation of these new RANNs.

The results of the current study are varied and complex; in the next few paragraphs we provide a broad synthesis of the most salient points of Tables 7–9, omitting any special mention of several significant brain–behavioral cross-domain correlations that appeared. Table 7 lists all findings pertaining to the classification accuracy of the reference label (MEM, FLUID, SPEED, VOCAB) in the validation sample of participants aged 31 and above. The first key finding was that the RANNs derived from the younger group identified the underlying reference ability from task related activation with great accuracy in the older subjects. Further, classification accuracy did not decline with age across participants. This suggests that unique networks associated with each reference ability as identified in young people are maintained with aging. This argues for a relative maintenance of specific neural network for each ability with age, and to some degree argues against the dedifferentiation hypothesis. In full disclosure, we emphasize that we presented one of many possible analytic frameworks, only using task design information for the derivation and validity test of our RANNs. To us, this strategy presented itself as the simplest and most obvious one to pursue. Of course, more elaborate strategies that simultaneously use behavioral performance constraints are conceivable too. The invariance across age observed for the RANNs derived in our 'minimalist' framework might not, and indeed is unlikely to, persist for these elaborate strategies since bringing more constraints to bear on the data is likely to boost particular sample dependencies. Thus, age-related changes might be found concerning both usage and topographic composition of networks that can classify the type of cognitive process *and* at the same time give a full account of behavioral performance.

In exploring the covariates that were associated with classification accuracy, we found that classification accuracy was reduced in people with lower brain volume, consistent with the idea that age-related brain changes may impact the integrity of these networks. Similarly, classification accuracy was related to verbal intelligence and the DRS in the same direction.

In addition, classification accuracy was associated with RANN network scores across participants in the validation sample, which was expected and demonstrated the consistency of our analysis framework. This was noted both for overall classification accuracy across all RANNs, and any specific reference ability's hit rate which was closely associated with its corresponding RANN pattern score. Interestingly, there were even associations between behavioral performance and classification accuracy, most notably for fluid reasoning, but also between overall classification accuracy and overall performance. This is remarkable since behavioral performance did not enter in any way into the derivation step of the RANN networks in the derivation sample.

Table 8
Bivariate relationships of RANN scores to demographic and performance variables. The RANN were derived from participants aged 20–30, while the relationships shown below are in participants age 31 and above. Correlation coefficients and p-levels are listed. Bolded cells indicate statistical significance at p ≤ 0.05. The VOCAB-RANN pattern score correlated inversely with VOCAB performance, but this inverse correlation was forced by an overly influential data point, whose removal voids statistical significance. All other reported significant correlations were robust.

	MEM-RANN	FLUID-RANN	SPEED-RANN	VOCAB-RANN
Age	R = 0.05;p = 0.42	R = −0.13;p = 0.06	R = −0.07;p = 0.32	R = −0.09;p = 0.16
Years of education	R = 0.06;p = 0.36	R = 0.09;p = 0.17	R = 0.13;p = 0.07	R = −0.09;p = 0.17
NARTIQ	R = 0.04;p = 0.55	R = −0.01; p = 0.85	R = 0.14;p = 0.06	**R = −0.20;p = 0.01**
DRS	R = 0.11;p = 0.13	R = 0.07; p = 0.35	**R = 0.15;p = 0.04**	R = −0.12;p = 0.08
Mean gray matter volume	**R = 0.14;p = 0.05**	**R = 0.17;p = 0.02**	R = 0.08;p = 0.25	R = 0.11;p = 0.11
Mean gray matter thickness	R = −0.03;p = 0.71	R = 0.01;p = 0.91	R = 0.06;p = 0.37	R = −0.01;p = 0.81
WMH	R = 0.001; p = 0.99	R = −0.06; p = 0.37	R = 0.04; p = 0.53	R = −0.03;p = 0.67
MEM performance	R = 0.06;p = 0.40	**R = 0.17;p = 0.01**	**R = 0.19;p = 0.01**	**R = 0.19;p = 0.008**
FLUID performance	R = 0.11;p = 0.11	**R = 0.22;p = 0.001**	R = 0.09;p = 0.21	R = 0.21;p = 0.09
SPEED performance	R = −0.17;p = 0.01	R = −0.07;p = 0.33	R = 0.07;p = 0.35	R = −0.07;p = 0.35
VOCAB performance	R = 0.06;p = 0.42	R = −0.06;p = 0.36	**R = 0.14;p = 0.05**	**R = 0.13;p = 0.05**

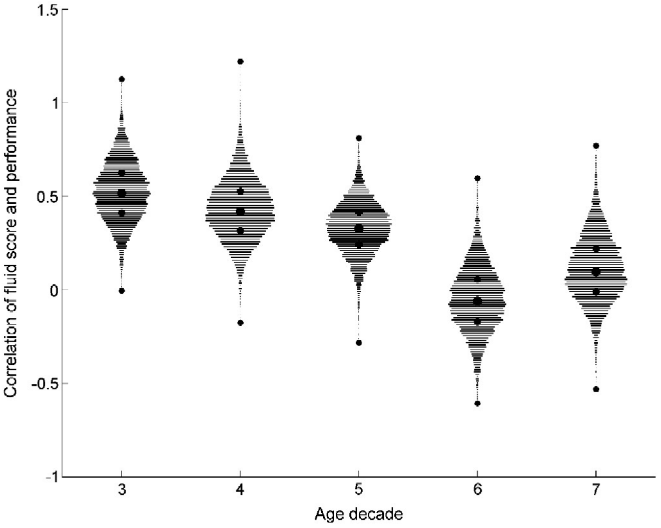

Fig. 4. 10,000 bootstrap samples of the brain-behavioral correlation (= Fisher-Z) between the FLUID-RANN network-score and FLUID performance, broken down by age decade. A decreasing linear trend can be observed. The one-tailed p-level obtained from the 10,000 samples is p = 0.0064.

In Table 8 we explored whether the degree of expression of each RANN was associated with demographic variables, neural measures and behavioral performance. The table shows no associations with age for any of the networks, but there are associations with mean brain volume and DRS score in a positive direction for some of the networks with one major exception: expression of the VOCAB-RANN correlated robustly with verbal intelligence in a *negative* direction. Since demographic information, similarly to behavioral performance, was absent in the network-derivation step, such a negative correlation is entirely possible: apparently, participants of superior verbal intelligence need to employ the VOCAB-RANN to a lesser degree.

We also probed for correlations between behavioral performance and RANN network expression. Similarly to the relationship with demographic information, it is possible that inter-individual differences in performance at any age could be related to network scores, even though performance information was not utilized for network derivation and network expression remained constant across age. Since Pearson correlation, against common belief, does not possess the mathematical property of transitivity, an association between network scores and behavioral performance cannot be excluded on account of different behavior of both variables with respect to age. It was therefore interesting to find that for at least one ability, fluid reasoning, there was a significant relationship between network expression and behavioral performance. In addition, the relationship between RANN pattern score and performance for MEM was at borderline significance. These observations hint at aspects of network expression that are responsible for inter-individual differences in performance. Further, we noted that the two RANNS whose expression was associated with behavioral performance were also sensitive to differential gray matter volume. The FLUID-RANN additionally showed an interaction with age in influencing behavioral performance: with increasing participant age, the association between FLUID-RANN scores and behavioral performance became weaker. This suggests that, while still appropriate for neurally based classification of the type of cognitive process the subject is engaged in, with advancing age the FLUID-RANN cannot give a satisfactory account of behavioral performance any longer: additional, possibly compensatory, components of brain activation might come into play. Despite the age-invariance of the FLUID-RANN itself, its emergent brain-behavior relationship is not age invariant.

In addition to the RANNs which best achieve discrimination of the reference abilities from each other, we also identified the neural substrate of common aspects of task processing. In addition to the specific reference abilities, shared variance across all tasks dwarfs the specific effects and needs to be taken into consideration as well, also since it might reveal meaningful associations with age. To do so, we derived one common pattern of activation present in all 12 tasks in the derivation sample of participants aged 20–30. Topographically, this pattern involved regions of activation and de-activation reminiscent of the task-positive network (Elton and Gao, 2015) and default-mode network (Raichle, 2015), respectively. In the validation sample (age 31–80), expression of this network during performance of any reference domain's tasks was not correlated with the respective behavioral performance, although it did show associations to age (Table 9): pattern expression was positively associated with age in the MEM and SPEED tasks (and marginally in the VOCAB tasks), while being negatively associated with age for the FLUID tasks.

This lack of invariance across age in the common pattern score though was complemented by the persistence across age of the specificity of pattern employment with respective to the reference domains: both derivation and validation samples showed convergent and

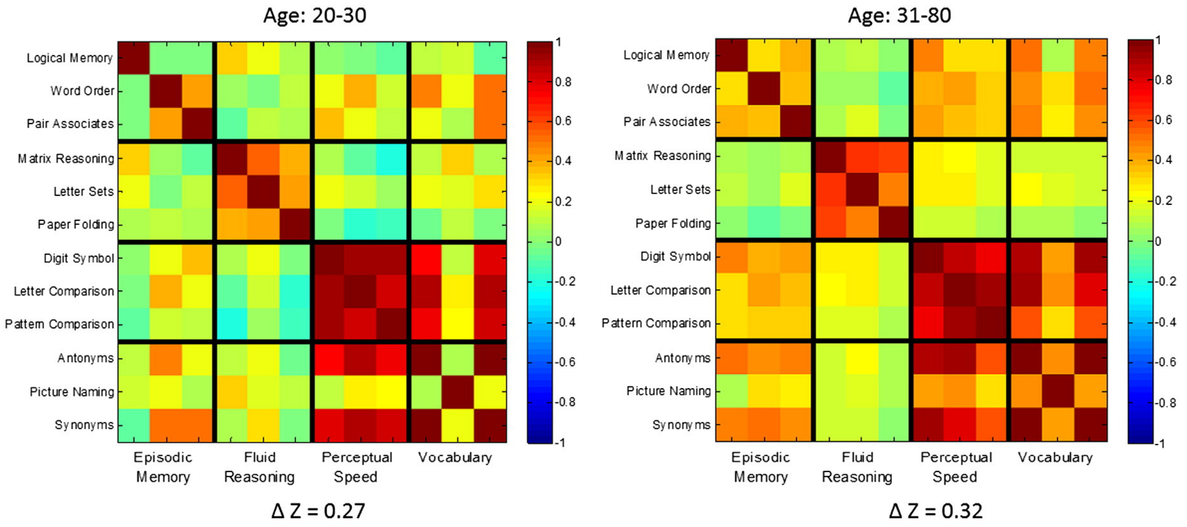

Fig. 5. Fisher-Z correlation matrices for the pattern score of the common activation in 12 tasks, displayed separately for the derivation sample of younger participants (ages 20–30, left panel)) and the validation sample (ages 31–80, right panel). Both groups show good overall convergent and discriminant validity ($p < 0.0001$), i.e. tasks belonging to the same reference domains show higher correlation than tasks belonging to different reference domains. One exception are the SPEED and VOCAB domains which lack good discriminant ability in both samples. A comparison of ΔZ between samples failed to show any significant difference (p = 0.77).

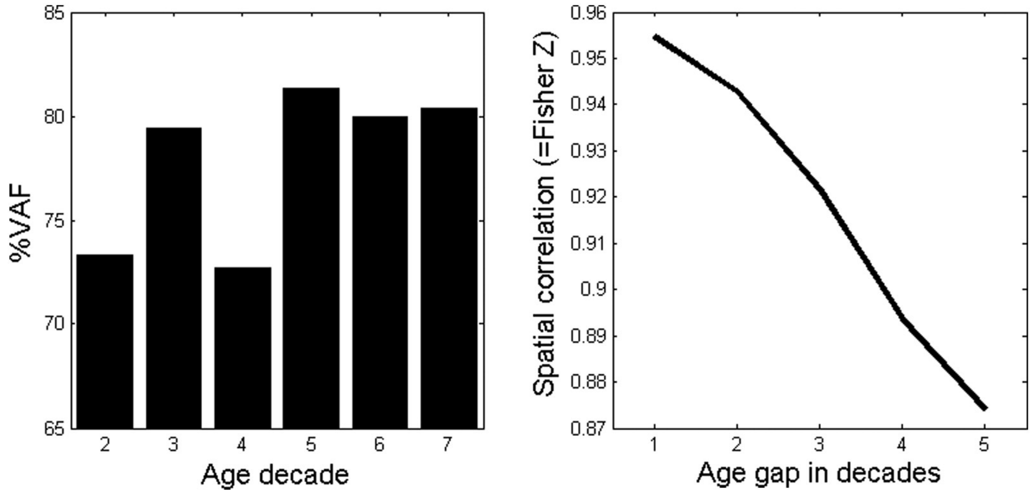

Fig. 6. The left panel shows the %VAF of the common activation pattern as a function of age decade. The right panel shows the average spatial correlation between the common patterns in different age decades as a function of the gap between the age decades. Permutation tests reveal that %VAF shows a significant *increasing* linear trend with age decade and that the spatial correlation shows a *decreasing* trend with the age gap.

discriminant validity of pattern usage with respect to the reference domains in that pairs of tasks chosen from the same reference domain in general had a higher pattern-score correlation than pairs of tasks chosen from *different* reference domains (with the exception of the lack of discriminant validity between SPEED and VOCAB domains). This construct validity did not manifest any difference between younger and older group, i.e. pattern usage did not evidence any ‘de-differentiation’ for the participants of higher age: their usage of the common activation pattern was as specific to the reference ability probed in the scanner as in their younger peers.

While the common activation pattern derived in young showed a similar specificity to cognitive domain in older participants, we found that when the common activation pattern was derived for each age decade separately the percentage variance accounted for by each age-specific common pattern increased significantly with the age decade, i.e. task processing aspects common to all tasks were more dominant than domain-specific aspects for older the participants.

Each of the four RANN patterns consists of a unique set of brain regions whose brain activation discriminate each ability from the others. Of the regions that show high loadings in the perceptual speed RA in Table 5, the left postcentral gyrus and the left inferior parietal cortex were significant in an automated meta-analysis over 114 studies conducted by Neurosynth for visuomotor tasks (Yarkoni et al., 2011), showing that these two regions were consistently reported by previous studies to be associated with visuomotor tasks. Bilateral activation was found in the dorsal striatum (Caudate and Putamen) which have an essential role in motor coordination (DeLong, 2000). The other regions, bilateral visual cortices and the left precental gyrus, represent important input and output processes. For the MEM-RANN, automated meta-analysis was also conducted on 270 studies using Neurosynth. Five of the regions with high positive loadings, including two of the highest z values (the left Calcarine and the left middle Occipital gyrus), were consistently reported by previous studies on episodic memory. Regions with high loadings in the right Precentral are involved in motor planning. Fluid reasoning RANN showed high loadings in the right Inferior Parietal area, which coincides with one of the regions from an automated meta-analysis over 142 studies conducted by Neurosynth for reasoning. Inferior Parietal lobule has been associated with Raven's matrix reasoning in normal controls (Yamada et al., 2012), which is also one of the fluid reasoning tasks administered in the current test battery. Lastly, for Vocabulary RANN, positive loadings were found in bilateral pre and post central gyri, parts of which were also significant in Neurosynth's automated meta-analysis of 152 studies for naming.

It is important to stress that only the regions that allow for maximal discrimination of one RA from the other 3 would load highly on the RANN pattern derived here. Regions that are common even among two of the RANNs would have low loadings in the patterns. Also, as mentioned in the Methods sections, our first-level modeling did not allow for any fine-grained separation of stimulus presentation and behavioral responses, so we have to rely on our group-level RANN derivation to capture only cognitive processes pertinent to the RA in question, with assignment of generic stimulus-presentation effects to the common activation pattern. This vagueness motivates inclusion of behavioral-performance information in future updates of our analytic framework.

Table 9
Bivariate relationships of scores of the common activation pattern for a particular reference domain to demographic and performance variables of the same reference domain. The common pattern was derived from participants aged 20–30, while the relationships shown below are in participants age 31 and above. Correlation coefficients and p-levels are listed. Bolded cells indicate statistical significance at p ≤ 0.05.

Common pattern score	MEM domain	FLUID domain	SPEED domain	VOCAB domain
Age	**R = 0.18; p = 0.01**	**R = −0.17; p = 0.03**	**R = 0.26; p = 0.0005**	R = 0.14; p = 0.06
Years of education	R = 0.07; p = 0.34	R = −0.03; p = 0.67	**R = 0.15; p = 0.05**	R = 0.04; p = 0.64
NARTIQ	R = −0.01; p = 0.92	R = −0.13; p = 0.13	R = 0.06; p = 0.45	R = −0.02; p = 0.76
DRS	R = −0.12; p = 0.14	R = −0.10; p = 0.23	**R = −0.18; p = 0.03**	R = −0.14; p = 0.08
Mean gray matter volume	R = 0.07; p = 0.38	R = 0.10; p = 0.22	R = 0.09; p = 0.26	R = 0.11; p = 0.15
Mean gray matter thickness	R = −0.11;p = 0.15	R = 0.13; p = 0.09	R = −0.11; p = 0.17	R = −0.06; p = 0.42
WMH	R = −0.02; p = 0.77	R = 0.04; p = 0.65	R = 0.01; p = 0.91	R = −0.02; p = 0.82
MEM performance	R = −0.01; p = 0.86	R = 0.008; p = 0.91	R = 0.04; p = 0.58	R = 0.03; p = 0.75
FLUID performance	R = 0.004; p = 0.95	R = −0.09; p = 0.25	R = −0.04; p = 0.58	R = −0.02; p = 0.81
SPEED performance	R = −0.10; p = 0.20	**R = −0.17; p = 0.03**	R = −0.07; p = 0.40	R = −0.05; p = 0.48
VOCAB performance	R = 0.06; p = 0.44	R = −0.14; p = 0.09	**R = 0.16; p = 0.04**	R = 0.01; p = 0.86

For example, hippocampus is not among the regions with the highest loadings in the MEM-pattern, suggesting that activation in the hippocampus was found in more than one RA, but not necessarily in *every* RA (which would guarantee capture in the common pattern). It is likely that hippocampal activation was found in matrix reasoning, one of the tasks for FLUID. (Pihlajamaki et al., 2004) reported hippocampal activation in a task involving presentation of objects in different spatial configurations as well as presentation of novel objects. Matrix reasoning shares some of the processes with this task due to the similarity in the nature of the task demand such as detection of objects in different spatial configurations across the matrix of cells.

While these findings strengthen our confidence that there is a specific RANN associated with each reference ability, we do not consider those derived here as our final representation of the RANNs. Our hope is that the final RANNs will not only meet the criteria of being specific to each reference ability, but that their expression will also be associated with task performance. Further, with enough data support, RANNs specific to each age decade can be derived, which allows a more thorough assessment of age-related changes in the topographic composition as well. In this vein, the current finding of age invariance in the classification performance has no implications for the age-relationships of topographic composition and brain–behavioral correlations of networks derived with an approach that integrates behavioral performance and uses the complete data set. As another note of caution, we cannot rule out the phenomenon of 'super normal' elders which, in contrast to the population at large, might have boosted the level of network scores for the elders in our sample: despite our efforts at avoiding recruitment biases, the group of participants above age 60 who successfully enrolled in the study ended up possessing higher verbal intelligence and more years education than its younger peer group (Table 1). Thus we cannot exclude the possibility of a healthy survivor effect in our older groups. The RANN scores did show associations with brain volume in several instances, although this extended to participants in middle age, without any age-related changes in this association. Also keeping in mind the cross-sectional study design, the association of RANN scores with brain volume thus cannot be unequivocally attributed to the effects of aging. Further, white-matter hyper-intensities, often taken as the clearest indication of aging in the brain, with a near perfect absence in younger people, were not associated with *any* of the findings resulting from our RANN analysis. This would argue against a possible distortion of findings by 'super normal' elders which artificially induce age-invariant effects.

In summary, current findings demonstrate that there are distinct neural networks underlying each of the four reference abilities, that these networks remain intact with aging, and that on an individual basis, these networks are specifically and appropriately recruited in response to the nature of the task. Our hope is that more definitive representations of these RANNs will put us in a position to better understand the neural processes that help maintain cognitive function with aging, and also the age-related neural or biological changes that may influence the expression of these networks with aging and result in poor performance in some elders.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2015.10.077.

Acknowledgments

The authors gratefully acknowledge funding support through grant NIH/NIA R01 AG038465 - Exploring Cognitive Aging Using Reference Ability Neural Networks (PI: Stern).

References

Blessed, G., Tomlinson, B.E., Roth, M., 1968. The association between quantitative measures of senile change in the cerebral grey matter of elderly subjects. Br. J. Psychol. 114, 797–811.

Brickman, A.M., Sneed, J.R., Provenzano, F.A., Garcon, E., Johnert, L., Muraskin, J., Yeung, L.K., Zimmerman, M.E., Roose, S.P., 2011. Quantitative approaches for assessment of white matter hyperintensities in elderly populations. Psychiatry Res. 193, 101–106.

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multimodel Inference. Springer Verlag, New York.

DeLong, M.R., 2000. Basal Ganglia. In: Kandel, E.R., Schwartz, J.H., Jessell, T.M. (Eds.), Principles of neural science. McGraw-Hill, Health Professions Division, New York (pp. xli, 1414 p).

Efron, B., Tibshirani, R.J., 1998. An Introduction to the Bootstrap. CRC Press, LLC, Boca Raton.

Ekstrom, R.B., French, J.W., Harman, H.H., Dermen, D., 1976. Manual for kit of factor-referenced cognitive tests. Princeton.

Elton, A., Gao, W., 2015. Task-positive Functional Connectivity of the Default Mode Network Transcends Task Domain. J. Cogn. Neurosci. 1–13.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355.

Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Segonne, F., Salat, D.H., Busa, E., Seidman, L.J., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B., Dale, A.M., 2004. Automatically Parcellating the Human Cerebral Cortex. Cereb. Cortex 14, 11–22.

Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The elements of statistical learning : data mining, inference, and prediction. 2nd ed. Springer, New York.

Mattis, S., 1988. Dementia Rating Scale (DRS). Psychological Assessment Resources, Odessa, FL.

McIntosh, A.R., Bookstein, F.L., Haxby, J.V., Grady, C.L., 1996. Spatial pattern analysis of functional brain images using partial least squares. Neuroimage 3, 143–157.

McIntosh, A.R., Lobaugh, N.J., 2004. Partial least squares analysis of neuroimaging data: applications and advances. Neuroimage 23 (Suppl. 1), S250–S263.

Pihlajamaki, M., Tanila, H., Kononen, M., Hanninen, T., Hamalainen, A., Soininen, H., Aronen, H.J., 2004. Visual presentation of novel objects and new spatial arrangements of objects differentially activates the medial temporal lobe subareas in humans. Eur. J. Neurosci. 19, 1939–1949.

Raichle, M.E., 2015. The Brain's Default Mode Network. Annu. Rev. Neurosci. 38, 433–447.

Raven, J.C., 1962. Advanced progressive matrices, set II. H.K. Lewis, London, UK.

Salthouse, T.A., 1993. Speed and knowledge as determinants of adult age differences in verbal tasks. J. Gerontol. 48, 29–36.

Salthouse, T.A., 1998. Independence of age-related influences on cognitive abilities across the life span. Dev. Psychol. 34, 851–864.

Salthouse, T.A., 2005. Relations between cognitive abilities and measures of executive functioning. Neuropsychology 19, 532–545.

Salthouse, T.A., 2009. Decomposing age correlations on neuropsychological and cognitive variables. J. Int. Neuropsychol. Soc. 15, 650–661.

Salthouse, T.A., Babcock, R.L., 1991. Decomposing adult age differences in working memory. Dev. Psychol. 27, 763–776.

Salthouse, T.A., Ferrer-Caja, E., 2003. What needs to be explained to account for age-related effects on multiple cognitive variables? Psychol. Aging 18, 91–110.

Salthouse, T.A., Pink, J.E., Tucker-Drob, E.M., 2008. Contextual analysis of fluid intelligence. Intelligence 36, 464–486.

Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., Niazy, R.K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J.M., Matthews, P.M., 2004. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23 (Suppl. 1), S208–S219.

Stern, Y., Habeck, C., Steffener, J., Barulli, D., Gazes, Y., Razlighi, Q., Shaked, D., Salthouse, T., 2014. The Reference Ability Neural Network Study: motivation, design, and initial feasibility analyses. Neuroimage 103, 139–151.

Woodcock, R.W., Johnson, M.B., Mather, N., 1989. Woodcock-Johnson Psycho-Educational Battery—Revised. DLM Teaching Resources.

Yamada, T., Ohta, H., Watanabe, H., Kanai, C., Tani, M., Ohno, T., Takayama, Y., Iwanami, A., Kato, N., Hashimoto, R., 2012. Functional alterations in neural substrates of geometric reasoning in adults with high-functioning autism. PLoS One 7, e43220.

Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., Wager, T.D., 2011. Large-scale automated synthesis of human functional neuroimaging data. Nat. Methods 8, 665–670.