the british psychological society
promoting excellence in psychology

Journal of Neuropsychology (2020)

www.wileyonlinelibrary.com

Reliable change in neuropsychological test scores is associated with brain atrophy in older adults

Shayne S.-H. Lin[1,2], Evan Fletcher[3], Brandon E. Gavett[1,4]* and for the Alzheimer's Disease Neuroimaging Initiative†

[1]Department of Psychology, University of Colorado at Colorado Springs, Colorado Springs, Colorado, USA
[2]Department of Psychology, The University of Alabama, Tuscaloosa, Alabama, USA
[3]Department of Neurology, University of California at Davis, California, USA
[4]School of Psychological Science, University of Western Australia, Perth, Western Australia, Australia

The reliable change index (RCI) is a commonly used method for interpreting change in neuropsychological test scores over time. However, the RCI is a psychometric method that, to date, has not been validated against neuroanatomical changes. Longitudinal neuroimaging and neuropsychological data from baseline and one-year follow-up visits were retrieved from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The RCI was used to identify participants showing reliable decline on memory (ADNI-Mem; $N = 450$) and executive functioning (ADNI-EF; $N = 456$) factor scores. For each factor score, two groups (reliable change vs. no reliable change) were matched on potential baseline confounding variables. Longitudinal neuroanatomical data were analysed using tensor-based morphometry. Analysis revealed that reliable change on ADNI-Mem was associated with atrophy in the medial temporal cortex, limbic cortex, temporal lobe and some regions of the parietal lobe. Similar atrophy patterns were found for reliable change on ADNI-EF, except that atrophy extended to the frontal lobe and the atrophy was more extensive and of higher magnitude. The current study not only validates clinical usage of the RCI with neuroanatomical evidence of associated underlying brain change but also suggests patterns of likely brain atrophy when reliable cognitive decline is detected.

Dementia is a prevalent disorder permeating the older adult population throughout the world. Over 35 million older adults are affected by dementia, and this number is estimated to surge to 65 million in 2030 and 115 million in 2050 (Prince *et al.*, 2013). The most common cause of dementia is Alzheimer's disease (AD), affecting approximately 33.9 million people worldwide and 5.3 million people in the United States (Brookmeyer, Johnson, Ziegler-Graham, & Arrighi, 2007). AD is described as causing an insidious decline in cognition, as it gradually affects the brain and, consequently, cognitive functioning. In

*Correspondence should be addressed to Brandon E. Gavett, School of Psychological Science, University of Western Australia (M304), 35 Stirling Highway, 6009 Perth, WA, Australia (email: brandon.gavett@uwa.edu.au).
†Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

DOI:10.1111/jnp.12226

particular, AD has demonstrated associations with patterns of brain atrophy that correspond to predictable changes in cognitive ability. The literature has found solid evidence to link episodic memory to circuitry involving the medial temporal lobes (e.g., hippocampus, entorhinal cortex), limbic system, thalamus and white matter pathways connecting these structures (e.g., Burianova, McIntosh, & Grady, 2010; Cummings, Tomiyasu, Read, & Benson, 1984; Danet *et al.*, 2015; Hamani *et al.*, 2008; Maddock, Garrett, & Buonocore, 2001; Rodrigue & Raz, 2004; Tsivilis *et al.*, 2008). Executive functions, on the other hand, have been more closely linked to frontal lobe functioning (Konishi *et al.*, 1998; Mentzel *et al.*, 1998; Volz *et al.*, 1997), prefrontal grey matter volume (Gunning-Dixon & Raz, 2003) and underlying white matter integrity (Kerchner *et al.*, 2012).

An essential clinical diagnostic feature of all neurodegenerative dementias, including AD, is the presence of documented cognitive change from a previous level (American Psychiatric Association, 2013; McKhann *et al.*, 2011). When neuropsychological test score changes are used to make inferences about underlying brain changes, test scores must demonstrate criterion validity for this purpose. Although numerous studies have demonstrated associations between brain change and cognitive change in group data (e.g., Fletcher *et al.*, 2018), there are a number of practical and conceptual issues that can complicate detection of cognitive change in individual patients.

Challenges in idiographic detection of cognitive change include, but are not limited to, low test–retest reliability, practice effects, floor and ceiling effects, and regression to the mean. In serial assessment, a test should produce consistent results between two time points. Since no test possesses perfect test–retest reliability – that is, a total lack of measurement error – score fluctuations over time might not reflect true changes but merely error of the test itself (Bowden & Finch, 2017). Tests with lower reliability are thus more susceptible to measurement error, which can either mask or exaggerate true change in the ability being measured.

Although in some cases it may be inadvisable to dichotomize continuous variables such as the magnitude of change in a test score, for individual clinical applications, questions often must be answered about whether the observed change is 'reliable'; that is, large enough to be reasonably certain that it is not due to measurement error. One of the methods commonly used in clinical settings to make decisions about the absence or presence of reliable change is the reliable change index (RCI). In its early form, the RCI considered only the initial test score and standard error when predicting the retest score (Christensen & Mendoza, 1986). Nevertheless, the formulae to compute RCI have evolved throughout the years, adding more information to account for confounding factors, such as the reliability of a measurement (Jacobson & Truax, 1991) and practice effects (Chelune, Naugle, Luders, Sedlak, & Awad, 1993). A more advanced version of the RCI, known as the standardized regression-based formula, was formulated using linear regression, which accounts for regression to the mean (McSweeny, Naugle, Chelune, & Luders, 1993).

Based on an examinee's baseline test score and the psychometric properties of the test, the RCI generates a confidence interval representing the range of retest scores that would be expected to occur with a given probability (here, 90%) simply due to measurement error (i.e., when no true change in ability occurs). Therefore, when an individual's retest score falls outside the range of the RCI interval, they are considered as having shown reliable (i.e., statistically significant) change. Such a psychometric approach offers clinicians an objective and quantifiable method to make clinical decisions about whether an observed test score change was or was not

produced simply due to chance under a certain level of confidence. Because this method is essentially equivalent to null hypothesis significance testing, however, it suffers from the same issue of statistical versus clinical significance that is often discussed in the clinical literature (Millis, 2003). Therefore, validation of the RCI against a meaningful criterion standard, such as documented brain change, is essential to demonstrating its criterion validity and clinical significance.

In one of the only studies known to examine the criterion validity of RCI methods against neuroimaging data, Duff, Suhrie, Dalley, Anderson and Hoffman (2019) evaluated reliable change on neuropsychological tests in a sample of 25 older adults over a one-week retest interval. Their results showed that a regression-based approach to calculating RCIs was associated with neuroimaging measures of hippocampal volume and amyloid deposition at baseline. This gave validation to regression-based formulas for RCI; however, their results were based on a small sample of one-time neuroimaging measurements and follow-up cognitive assessment over a relatively short retest interval. We propose here to extend that work by testing the associations of RCI to longitudinal brain tissue volume change, using an exploratory voxel-based signature region approach (Bakkour *et al*., 2009; Dickerson *et al*., 2009; Fletcher *et al*., 2013; Hua *et al*., 2008) for computations of local one-year brain atrophy rates. To build on the work of Duff *et al*. (2019) and further validate the regression-based RCI against neuroanatomical evidence, the current study used comprehensive and longitudinal brain structural measures, a one-year retest interval, a large and cognitively heterogeneous sample, and a matching procedure to control for the influence of numerous potential confounding variables.

The present study

To summarize, the present study aimed to examine the criterion validity of the RCI, applied to two composite cognitive scores from ADNI – ADNI-Mem (memory) and ADNI-EF (executive function) – for its associations with longitudinal brain atrophy across a period of one year. To mirror the dichotomous decisions about change vs. no change often faced by clinicians, the RCI was used for each factor score separately to identify one group of participants showing *reliable change* (RC) and a matched group of participants showing *no reliable change* (NC). We hypothesized that, compared to the NC groups, the RC groups would show greater brain atrophy in a manner that corresponds to the neuroanatomical regions thought to underlie performance on tests of episodic memory and executive functioning. That is, reliable change on ADNI-Mem is expected to correspond to greater atrophy in temporolimbic regions, whereas reliable change on ADNI-EF is expected to correspond to greater atrophy in frontal-striatal regions.

Method

Participants

Data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), other biological markers, and clinical and neuropsychological assessment can be

combined to measure the progression of mild cognitive impairment (MCI) and early AD. (For up-to-date information, see www.adni-info.org.) Therefore, all data used were archival. The current research protocol was evaluated by The University of Colorado Colorado Springs Institutional Review Board, who determined that the current research was exempt from review because it was not human subjects research. Inclusion criteria for the present study were that participants should complete both neuropsychological assessment and MRI scanning at two time points approximately one year apart.

Materials

ADNI-Mem and ADNI-EF

ADNI-Mem (Crane *et al.*, 2012) and ADNI-EF (Gibbons *et al.*, 2012), which are psychometrically sophisticated composite scores of memory and executive functioning, respectively, were used to identify individuals who experienced reliable change in cognition. Using confirmatory factor analysis, ADNI-Mem factor scores are derived from observed scores on the Mini-Mental State Examination (Folstein, Folstein, & McHugh, 1975), AD Assessment Schedule – Cognition (Rosen, Mohs, & Davis, 1984), Auditory Verbal Learning Test (Rey, 1941) and Logical Memory (Wechsler, 1987). Similarly, ADNI-EF factor scores are derived from observed scores on Digit Symbol Coding and Digit Span (Wechsler, 1981), Trail Making Test A and B (Reitan and Wolfson, 1993; Strauss, Sherman, & Spreen, 2006), category fluency (Animal and Vegetable; Thurstone, 1938) and the clock drawing test (Kaplan, 1988). Both ADNI-Mem and ADNI-EF factor scores were scaled as z-scores ($M = 0$, $SD = 1$). The psychometric properties (i.e., test–retest reliability, means and standard deviations) of these factor scores used in the construction of reliable change indices were derived from the cognitively normal group in ADNI and are reported below.

MRI measures

Brain volume measurements were based on T1-weighted MRI scans acquired with 1.5- or 3-Tesla scanners. The data collected in the current study spans from ADNI1 to ADNI3 and the MRI protocol has changed throughout the years. As described below, we used matched samples to ensure that groups did not differ in MRI collection protocols. Baseline MRI scans and longitudinal data were processed with the University of California (UC) at Davis IDeA (Imaging of Dementia and Aging) laboratory's in-house pipeline (Fletcher *et al.*, 2013; Fletcher, Singh, Harvey, Carmichael, & DeCarli, 2012). Tensor-based morphometry (TBM) was conducted to analyse the MRI change data (Fletcher *et al.*, 2013). Briefly, for each individual subject, sequential scans were first linearly aligned to a common 'halfway' space. Then, a nonlinear deformation using TBM was calculated to capture local nonlinear variations between the two scans (Fletcher *et al.*, 2018). The determinant of the 3×3 Jacobian matrix of the deformation field yielded a local volume change factor at each voxel, which was then log-transformed (the log-Jacobian) to produce a symmetric distribution about zero, with negative values indicating tissue atrophy and positive values indicating volume expansion. The log-Jacobian provides approximate percentage volume change at each voxel. Both voxelwise and region of interest (ROI) analyses were performed using the log-Jacobian atrophy estimates; predefined ROIs were determined using the Mindboggle atlas of cortical parcellations (https://mindboggle.info/; Klein *et al.*, 2017).

Procedure

The present research includes two studies, the first for ADNI-Mem and the second for ADNI-EF. The procedures for these two studies were identical, with the exception of the cognitive test score used to identify reliable change. Participants who showed reliable cognitive decline on ADNI-Mem or ADNI-EF, using the methods described below, were selected from the data set and assigned to the RC group. Participants who did not show reliable change and who were matched on a number of background variables, described below, were assigned to the NC group.

Reliable change

To divide participants into RC and NC groups, McSweeny et al. (1993) RCI formula was implemented, which uses a simple regression-based approach. For each participant, a 90% RCI confidence interval for a predicted follow-up score was calculated based on the tests' psychometric and statistical properties and participants' baseline cognitive test scores. If the difference between the participants' observed and predicted follow-up scores fell below the 90% reliable change interval, the participant was assigned to the RC group. Because we predicted a directional effect (i.e., decline), we used a 1-tailed reliable change interval; as such, all participants not meeting our criteria for reliable change – even those who may have showed reliable improvement in their cognitive test scores over time – were considered to have experienced 'no change'.

Participant matching

To ensure that participants from the RC group and the NC group were equivalent on relevant confounding variables, a genetic matching procedure was applied to create two equivalent groups of equal sample size. Genetic matching is a statistical procedure that is used to generate groups of individuals that are comparable across a number of variables (Diamond & Sekhon, 2013). The confounding variables that were matched included baseline clinical diagnosis (cognitively normal, MCI or AD); demographic variables, including age, gender, years of education, race and ethnicity; baseline cognitive performance (explicated in the following sections); study-specific variables, including participants' baseline and follow-up data collection protocols (ADNI-1, ADNI-2, ADNI-GO, ADNI-3); the duration of the test–retest interval; and baseline whole brain volume, adjusted for intracranial volume.

Because the number of RC participants was smaller than the number of NC participants for both ADNI-Mem and ADNI-EF, the size of the RC group was used as the default sample size for matching purposes. The matching procedure used in the current research comes from the *Matching* package version 4.9-3 (Sekhon, 2011), an R (R Core Team, 2019) library that performs matching using a genetic algorithm. Details about this package have been explicated in Diamond and Sekhon (2013).

Data analysis

Neuroanatomical data were analysed with two techniques: voxelwise whole brain analysis and ROI analysis. In order to perform voxelwise analysis, all log-Jacobian atrophy maps in subject native space were nonlinearly warped to a common age-appropriate template brain (Kochunov *et al.*, 2001) using a cubic B-spline diffeomorphism (Rueckert, Aljabar, Heckemann, Hajnal, & Hammers, 2006). In each group, we performed an

exploratory, voxelwise test of the entire brain parenchyma, leading to a 'signature ROI' depiction of areas of the brain most strongly associated to cognitive outcome (Bakkour *et al.*, 2009; Dickerson *et al.*, 2009; Fletcher *et al.*, 2013; Hua *et al.*, 2008). These could then be compared for RC and NC groups to reveal differing patterns of brain atrophy most associated with each group. The signature ROIs were created by nonparametric superthreshold cluster testing, which has been found to produce similar results to the familywise error methods used by Statistical Parametric Mapping (SPM) software (Nichols & Holmes, 2002). This analysis was aimed at examining different patterns of brain atrophy for RC and NC groups relative to each of the cognitive domains. An alpha level of .001 was adopted with 1000 iterations of random permutations. For ROI analysis, mean log-Jacobian atrophy rates were compared between the RC and NC groups for brain grey matter regions parcellated by the Mindboggle atlas (Klein *et al.*, 2017).

Results

Study 1 - ADNI-Mem

As described above, we used the performance of participants clinically diagnosed as cognitively normal by ADNI at both baseline and follow-up (n = 450) to determine the psychometric properties of ADNI-Mem scores, so we could construct the RCI for this measure. The formulas used to determine reliable change are shown below.

$$y' = x_i \frac{r_{xy} s_y}{s_x} + \left(\overline{y} - \overline{x} \frac{r_{xy} s_y}{s_x} \right)$$

$$SE_E = s_y \sqrt{1 - r_{xy}^2}$$

where x_i is an individual participant's baseline ADNI-Mem score, r_{xy} is the test–retest reliability (r_{xy} = .79, 95% CI [.75, .82]), $\overline{x}$ and s_x are the mean ($\overline{x} = 1.04$) and standard deviation ($s_x = 0.57$) of ADNI-Mem scores at baseline, $\overline{y}$ and s_y are the mean ($\overline{y} = 1.12$) and standard deviation ($s_y = 0.62$) of ADNI-Mem scores at follow-up, y' is the predicted follow-up ADNI-Mem score, and SE_E is the standard error of the estimate. If the difference between an individual participant's observed follow-up ADNI-Mem score and their predicted follow-up ADNI-Mem score ($y_i - y'$) was lower than $-1.645 \times SE_E$, then the participant was classified as having shown reliable change (decline) on ADNI-Mem. This approach generates a unique y' value for each participant – depending on their observed score at baseline – and a single SE_E value for all participants. For ADNI-Mem, SE_E was .38, meaning that an individual's follow-up score needed to be more than .625 points lower than predicted to be considered 'reliable' change with 90% confidence.

Participant matching

From a total of 1412 ADNI participants with at least 2 visits and no missing data on the variables needed for matching, 225 participants were identified as exhibiting reliable change over one year on the ADNI-Mem factor score, using the McSweeney method (McSweeny et al., 1993) for identifying reliable decline applied to the psychometric data described above. A second sample of 225 participants was matched to this sample, with the exception that the second sample did not show reliable change on ADNI-Mem scores.

The success of the matching procedure in selecting 225 matched NC participants was judged using bootstrap p-values and effect sizes (e.g., standardized mean difference). Results from the genetic matching procedure revealed that the two groups were not statistically different from each other on any of the confounding variables (Table 1). Of these 450 participants, 44 ($n_{RC} = 20$, $n_{NC} = 24$) were in the sample of 450 individuals described in the previous section whose data were used to generate the sample means, standard deviations and test–retest reliability data used to identify reliable change.

Voxelwise exploratory brain analysis

After matching, whole brain analysis was done to understand voxelwise patterns of atrophy that differed between the two groups. Longitudinally, voxelwise whole brain analysis revealed that, whereas the RC group showed extensive bilateral atrophy, principally in the medial and lateral temporal lobes, the NC group displayed only a negligible amount of atrophy (Figure 1).

ROI analysis

Group ROI comparisons are shown in Table 2. These results show that the RC group experienced more atrophy than the NC group in supramarginal gyrus (219% more atrophy), inferior parietal lobule (135% more atrophy), superior temporal gyrus (56% more atrophy), inferior temporal gyrus (45% more atrophy), rostral anterior cingulate (44% more atrophy), entorhinal cortex (43% more atrophy), lateral temporal cortex (43% more atrophy), middle temporal gyrus (40% more atrophy), fusiform gyrus (33% more atrophy) and parahippocampal gyrus (20% more atrophy).

Study 2 – ADNI-EF

The performance of participants clinically diagnosed as cognitively normal by ADNI at both baseline and follow-up was used for the purposes of determining the psychometric properties of ADNI-EF scores. In this sample ($n = 421$), the test–retest reliability of ADNI-EF was $r_{xy} = .71$, 95% CI [.66, .76], average performance at baseline was $\overline{x} = 0.81$ ($s_x = 0.78$), and average performance at follow-up was $\overline{y} = 0.90$ ($s_y = 0.79$). The equations described in the ADNI-Mem section above were also applied to the ADNI-EF data to determine reliable change (decline) on this scale. For ADNI-EF, SE_E was 0.56, meaning that an individual's follow-up score needed to be more than 0.915 points below predicted to be considered 'reliable' change with 90% confidence.

Participant matching

From a total of 1412 ADNI participants with at least two visits and no missing data on the variables needed for matching, 228 participants were identified as exhibiting reliable decline over one year on the ADNI-EF factor scores. A second sample of 228 participants was matched to this sample, with the exception that the second sample did not show reliable change on ADNI-EF scores. Results from the genetic matching procedure revealed that the two groups were not statistically different from each other on any of the confounding variables, except for age (Table 3). Although the NC group was older than the RC group, the difference was small (1.56 years; $d = .213$). Of these 456 participants, 35 ($n_{RC} = 15$, $n_{NC} = 20$) were in the sample of 421 individuals described in the previous

Table 1. Participant demographics in study 1 – ADNI-Mem

Variable	No-change group	Reliable change group	*p*	SMD
N	225	225		
Baseline diagnosis			.990	
Cognitively normal	18	18		
Subjective cognitive impairment	3	3		
Early mild cognitive impairment	21	24		
Late mild cognitive impairment	94	90		
Alzheimer's disease	89	90		
Age (years)	$M = 74.70$ ($SD = 6.51$)	$M = 73.97$ ($SD = 7.37$)	.262	0.106
Gender	Male = 126; Female = 99	Male = 123; Female = 102	.850	
Education (years)	$M = 15.50$ ($SD = 2.75$)	$M = 15.46$ ($SD = 2.83$)	.866	0.016
Ethnicity			1.000	
Hispanic/ Latino	8	8		
Not Hispanic/ Latino	216	216		
Unknown	1	1		
Race			1.000	
Asian	6	1		
Black	5	9		
White	212	214		
Multiple	2	1		
CDR-SB baseline	$M = 2.75$ ($SD = 1.84$)	$M = 2.76$ ($SD = 1.93$)	.960	0.005
ADAS 11 baseline	$M = 14.43$ ($SD = 6.55$)	$M = 14.93$ ($SD = 6.83$)	.428	0.075
MMSE baseline	$M = 25.83$ ($SD = 2.77$)	$M = 25.48$ ($SD = 2.93$)	.198	0.122
LM delayed baseline	$M = 3.76$ ($SD = 4.64$)	$M = 3.84$ ($SD = 4.24$)	.865	0.016
TMT-B baseline	$M = 162.03$ ($SD = 88.04$)	$M = 168.82$ ($SD = 89.89$)	.419	0.076
FAQ baseline	$M = 7.51$ ($SD = 7.06$)	$M = 7.79$ ($SD = 7.46$)	.683	0.039
ADNI-Mem baseline	$M = -0.35$ ($SD = 0.80$)	$M = -0.33$ ($SD = 0.83$)	.744	0.031
ADNI-Mem follow-up	$M = -0.37$ ($SD = 0.83$)	$M = -0.94$ ($SD = 0.81$)	<.001	0.698
Follow-up interval (months)	$M = 12.20$ ($SD = 0.99$)	$M = 12.23$ ($SD = 1.22$)	.735	0.032

Continued

Table 1. (*Continued*)

Variable	No-change group	Reliable change group	*p*	SMD
ADNI baseline data sources			1.000	
ADNI-1	127	126		
ADNI-GO	3	7		
ADNI-2	95	92		
ADNI follow-up data sources			1.000	
ADNI-1	127	126		
ADNI-GO	0	0		
ADNI-2	98	99		
Adjusted whole brain volume	$M = -0.39$ ($SD = 0.88$)	$M = -0.39$ ($SD = 0.88$)	.985	0.002

Note. All values are from the baseline visit unless otherwise noted. ADNI = Alzheimer's disease neuroimaging initiative; SMD = standardized mean difference; CDR-SB = Clinical dementia rating–sum of boxes; ADAS = Alzheimer's disease assessment schedule; MMSE = Mini-mental state examination; LM = Logical memory test; TMT = Trail making test; FAQ = Functional activities questionnaire.

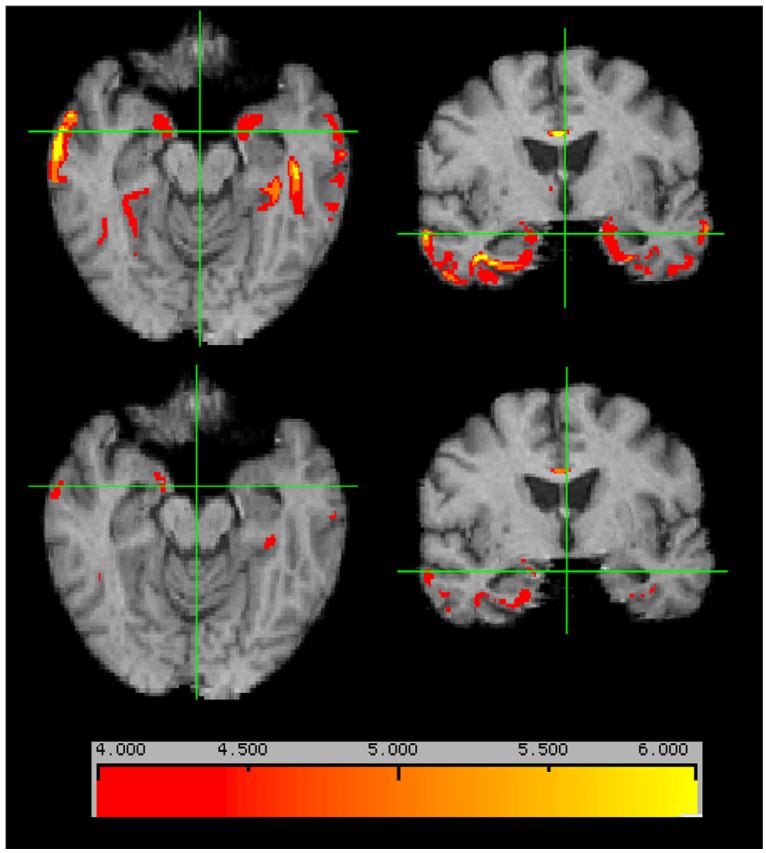

Figure 1. Significant clusters ($p < .001$) of tissue atrophy over one-year period for the RC group (top) and the NC group (bottom) in the ADNI-Mem study. Colour bar shows approximate percentage of volume loss at each voxel (corresponding to the log-Jacobian values from Table 2 multiplied by 100). Left column: axial views; right column: coronal views. Left side of each image (axial and coronal) = left hemisphere.

section whose data were used to generate the sample means, standard deviations and test–retest reliability data used to identify reliable change.

Voxelwise exploratory brain analysis

After matching, whole brain analysis was done to understand voxelwise patterns of atrophy that differed between the two groups. Longitudinally, voxelwise whole brain analysis revealed that, whereas the RC group showed extensive bilateral atrophy, principally in medial and lateral portions of the temporal lobes, the thalamus, corpus callosum, posterior cingulate and the medial orbitofrontal portions of the frontal lobe, the NC group displayed a smaller extent of atrophy with lower t magnitudes of significance (Figure 2). The atrophy pattern in the temporal lobe and the limbic cortex was similar in both ADNI-Mem and ADNI-EF. However, the atrophy pattern in the RC group in the ADNI-EF study was more extensive, and of overall higher signal magnitude than for the ADNI-Mem RC.

Table 2. Log-Jacobians from region of interest analysis for reliable change and no change on ADNI-Mem over a one-year period

Region of interest	Reliable change group; *M* (*SD*)	No-Change group; *M* (*SD*)	*t*	*p*	Cohen's *d*	Mean difference	95% Confidence interval for the mean difference
Medial temporal cortex	−.025 (.021)	−.019 (.021)	3.224	.001	−0.305	−.006	[−.010, −.002]
Temporal lobe	−.024 (.022)	−.017 (.024)	3.171	.002	−0.300	−.007	[−.011, −.003]
Fusiform gyrus	−.024 (.022)	−.018 (.022)	3.096	.002	−0.293	−.006	[−.010, −.002]
Lateral temporal cortex	−.023 (.025)	−.017 (.022)	2.936	.003	−0.278	−.007	[−.012, −.002]
Middle temporal gyrus	−.028 (.029)	−.020 (.030)	2.783	.006	−0.263	−.008	[−.013, −.002]
Inferior temporal gyrus	−.032 (.035)	−.022 (.038)	2.637	.009	−0.249	−.009	[−.002, −.016]
Superior temporal gyrus	−.014 (.021)	−.009 (.022)	2.621	.009	−0.248	−.005	[−.009, −.001]
Parahippocampal gyrus	−.024 (.020)	−.020 (.021)	2.303	.022	−0.218	−.004	[−.008, −.001]
Entorhinal cortex	−.030 (.038)	−.021 (.042)	2.288	.023	−0.216	−.009	[−.016, −.001]
Rostral anterior cingulate cortex	−.013 (.020)	−.009 (.020)	1.921	.055	−0.182	−.004	[−.007, .000]
Inferior parietal lobule	−.009 (.029)	−.004 (.031)	1.908	.057	−0.180	−.005	[−.011, .000]
Supramarginal gyrus	−.007 (.026)	−.002 (.026)	1.891	.059	−0.179	−.005	[−.009, .000]
Pars triangularis	−.005 (.029)	−.002 (.028)	1.405	.161	−0.133	−.004	[−.009, .002]
Rostral middle frontal gyrus	−.004 (.040)	.001 (.037)	1.391	.165	−0.132	−.005	[−.012, .002]
Insula	−.016 (.017)	−.014 (.016)	1.390	.165	−0.131	−.002	[−.005, .001]
Posterior cingulate Cortex	−.022 (.029)	−.018 (.029)	1.388	.166	−0.131	−.004	[−.009, .002]
Caudal anterior cingulate cortex	−.017 (.026)	−.014 (.023)	1.304	.193	−0.123	−.003	[−.008, .002]
Lateral occipital sulcus	−.008 (.027)	−.004 (.031)	1.308	.191	−0.124	−.004	[−.009, .002]
Hippocampus	.034 (.050)	.029 (.049)	1.183	.237	0.112	.006	[−.004, .015]
Paracentral lobule	.000 (.026)	−.002 (.023)	1.135	.257	0.107	.003	[−.002, .007]
Pars opercularis	−.008 (.022)	−.005 (.021)	1.122	.263	−0.106	−.002	[−.006, .002]
Isthmus of the Cingulate	−.021 (.021)	−.019 (.023)	1.099	.273	−0.104	−.002	[−.006, .002]
Transverse temporal gyrus	−.014 (.023)	−.011 (.024)	1.026	.305	-0.097	−.002	[−.007, .002]
Pars orbitalis	−.005 (.043)	−.001 (.051)	0.930	.353	−0.088	−.004	[−.013, .005]
Superior frontal gyrus	−.003 (.037)	.000 (.032)	0.871	.384	−0.082	−.003	[−.009, .004]
Lateral orbitofrontal cortex	−.010 (.037)	−.008 (.042)	0.758	.449	−0.072	−.003	[−.010, .005]
Lingual gyrus	−.004 (.013)	−.003 (.015)	0.756	.450	-0.072	−.001	[−.004, .002]

Continued

Table 2. (*Continued*)

Region of interest	Reliable change group; *M* (*SD*)	No-Change group; *M* (*SD*)	*t*	*p*	Cohen's *d*	Mean difference	95% Confidence interval for the mean difference
Superior parietal lobule	.004 (.040)	.002 (.032)	0.705	.481	0.067	.002	[−.004, .009]
Precentral gyrus	.002 (.027)	.000 (.025)	0.532	.595	0.050	.001	[−.004, .006]
Medial orbitofrontal cortex	−.010 (.031)	−.008 (.035)	0.528	.598	−0.050	−.002	[−.008, .004]
Precuneus	−.006 (.022)	−.007 (.019)	0.444	.657	0.042	.001	[−.003, .005]
Post-central gyrus	.010 (.033)	.008 (.029)	0.415	.678	0.039	.001	[−.005, .007]
Calcarine gyrus	.006 (.017)	.005 (.023)	0.096	.924	−0.009	−.000	[−.004, .004]

Table 3. Participant demographics in study 2 – ADNI-EF

Variable	No-Change Group	Reliable Change Group	*p*	SMD
N	228	228		
Baseline diagnosis			.619	
Cognitively normal	18	17		
Subjective cognitive Impairment	0	2		
Early mild cognitive impairment	22	22		
Late Mild cognitive impairment	90	82		
Alzheimer's Disease	98	105		
Age (years)	*M* = 75.10 (*SD* = 7.18)	*M* = 73.54 (*SD* = 7.53)	.024	0.213
Gender	M = 119; F = 109	M = 120; F = 108	1.000	
Education (years)	*M* = 15.39 (*SD* = 2.82)	*M* = 15.36 (*SD* = 3.14)	.913	0.010
Ethnicity			.785	
Hispanic/ Latino	3	4		
Not Hispanic/ Latino	224	222		
Unknown	1	2		
Race			1.000	
American Indian/Alaska Native	1	0		
Asian	2	1		
Black	10	12		
White	211	215		
Multiple	4	0		
CDR-SB baseline	*M* = 2.79 (*SD* = 1.87)	*M* = 2.93 (*SD* = 2.00)	.439	0.073
ADAS 11 baseline	*M* = 14.74 (*SD* = 5.78)	*M* = 15.87 (*SD* = 7.28)	.068	0.171
MMSE baseline	*M* = 25.65 (*SD* = 2.69)	*M* = 25.27 (*SD* = 2.91)	.147	0.136
LM delayed baseline	*M* = 3.65 (*SD* = 4.25)	*M* = 3.49 (*SD* = 4.20)	.690	0.037
TMT-B Baseline	*M* = 168.51 (*SD* = 89.76)	*M* = 178.82 (*SD* = 89.85)	.221	0.115
FAQ baseline	*M* = 7.54 (*SD* = 7.36)	*M* = 8.41 (*SD* = 7.55)	.214	0.117
ADNI-EF baseline	*M* = −0.39 (*SD* = 0.99)	*M* = −0.53 (*SD* = 1.09)	.149	0.135
ADNI-EF follow-up	*M* = −0.28 (*SD* = 0.97)	*M* = −1.36 (*SD* = 0.89)	< .001	1.160

Continued

Table 3. (*Continued*)

Variable	No-Change Group	Reliable Change Group	*p*	SMD
Follow-up Interval (months)	$M = 12.10$ ($SD = 0.50$)	$M = 12.15$ ($SD = 0.58$)	.293	0.099
ADNI baseline data sources			1.000	
ADNI-1	136	129		
ADNI-GO	8	9		
ADNI-2	84	90		
ADNI follow-up data sources			1.000	
ADNI-1	136	129		
ADNI-GO	1	0		
ADNI-2	91	99		
Adjusted whole brain volume	$M = -0.38$ ($SD = 0.93$)	$M = -0.44$ ($SD = 0.95$)	.524	0.060

Note. All values are from the baseline visit unless otherwise noted. SMD = standardized mean difference; ADNI = Alzheimer's disease neuroimaging initiative; CDR-SB = Clinical dementia rating–sum of boxes; ADAS = Alzheimer's disease assessment schedule; MMSE = Mini-mental state examination; LM = Logical memory test; TMT = Trail making test; FAQ = Functional activities questionnaire.

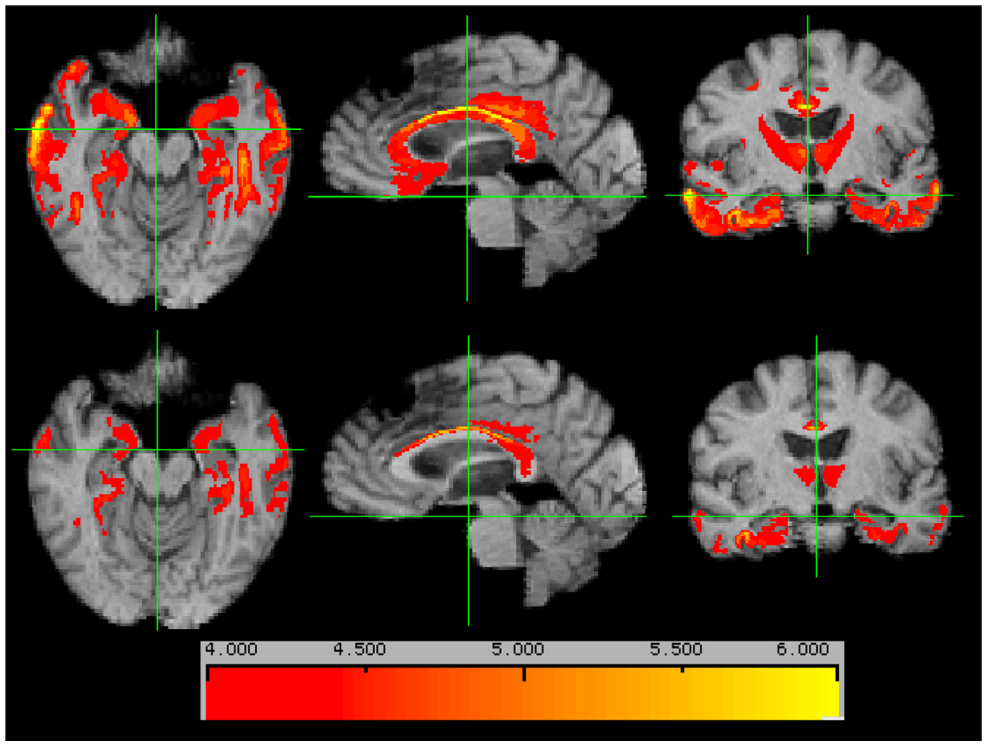

Figure 2. Significant clusters ($p < .001$) of tissue atrophy over one-year period for the RC group (top) and the NC group (bottom) in the ADNI-EF study. Colour bar shows approximate percentage of volume loss at each voxel (corresponding to the log-Jacobian values from Table 4 multiplied by 100). Left column: axial views; middle column: sagittal views; right column: coronal views. Left sides of the axial and coronal images = left hemisphere.

ROI analysis

Group ROI comparisons are shown in Table 4. These results show that the RC group experienced more atrophy than the NC group in precentral gyrus (>1000% more atrophy), rostral middle frontal gyrus (>1000% more atrophy), pars orbitalis (458% more atrophy), superior frontal gyrus (330% more atrophy), lateral orbitofrontal cortex (187% more atrophy), pars triangularis (179% more atrophy), pars opercularis (95% more atrophy), supramarginal gyrus (89% more atrophy), medial orbitofrontal sulcus (85% more atrophy), precuneus (62% more atrophy), superior temporal gyrus (60% more atrophy), hippocampus (59% less enlargement), entorhinal cortex (56% more atrophy), lateral temporal cortex (55% more atrophy), middle temporal cortex (50% more atrophy), inferior temporal lobule (47% more atrophy), posterior cingulate (42% more atrophy), fusiform gyrus (34% more atrophy), caudal anterior cingulate cortex (34% more atrophy), medial temporal cortex (33% more atrophy) and isthmus of the cingulate (25% more atrophy). It should also be noted that some of these extremely high values for percent change in atrophy were largely driven by very small values in the denominator, which represent the amount of atrophy (often not statistically different from 0) in the NC group. Therefore, the Cohen's *d* values reported in Tables 2 and 4 likely provide a more realistic measure of effect size than per cent change in atrophy rates. According to a common

Table 4. Log-Jacobians from region of interest analysis for reliable change and no change on ADNI-EF over a one-year period

Region of interest	Reliable change group	No-Change Group	*t*	*p*	Cohen's *d*	Mean difference	95% confidence interval for the mean difference
Middle temporal gyrus	−.030 (.030)	−.020 (.028)	3.734	<.001	−0.350	−.010	[−.016, −.005]
Temporal lobe	−.024 (.023)	−.017 (.022)	3.608	<.001	−0.339	−.0078	[−.012, −.003]
Hippocampus	.043 (.051)	.027 (.047)	−3.595	<.001	0.337	.016	[−.025, −.007]
Lateral temporal cortex	−.024 (.026)	−.016 (.023)	3.486	.001	−0.327	−.008	[−.013, −.004]
Medial temporal cortex	−.024 (.021)	−.018 (.021)	3.264	.001	−0.306	−.006	[−.010, −.003]
Pars opercularis	−.012 (.019)	−.006 (.019)	3.107	.002	−0.292	−.005	[−.009, −.002]
Fusiform gyrus	−.024 (.022)	−.018 (.021)	3.098	.002	−0.291	−.006	[−.010, −.002]
Superior temporal gyrus	−.016 (.021)	−.010 (.019)	3.052	.002	−0.286	−.006	[−.010, −.002]
Inferior temporal gyrus	−.031 (.040)	−.021 (.033)	2.867	.004	−0.269	−.010	[−.017, −.003]
Posterior cingulate cortex	−.027 (.031)	−.019 (.030)	2.828	.005	−0.265	−.008	[−.014, .002]
Entorhinal cortex	−.028 (.040)	−.018 (.040)	2.749	.006	−0.258	−.010	[−.018, −.003]
Isthmus	−.024 (.019)	−.019 (.020)	2.638	.009	−0.248	−.005	[−.009, −.001]
Rostral middle frontal gyrus	−.008 (.032)	.000 (.029)	2.624	.009	−0.246	−.008	[−.013, −.002]
Pars triangularis	−.009 (.025)	−.003 (.024)	2.573	.010	−0.242	−.006	[−.010, −.001]
Supramarginal gyrus	−.011 (.025)	−.006 (.022)	2.524	.012	−0.237	−.006	[−.010, −.001]
Superior frontal gyrus	−.009 (.034)	−.002 (.026)	2.449	.015	−0.230	−.007	[−.013, −.001]
Precentral gyrus	−.005 (.024)	−.000 (.021)	2.356	.019	−0.221	−.005	[−.009, −.001]
Caudal anterior cingulate cortex	−.020 (.024)	−.015 (.023)	2.334	.020	−0.219	−.005	[−.009, −.001]
Lateral orbitofrontal cortex	−.012 (.038)	−.004 (.031)	2.239	.026	−0.210	−.007	[−.014, −.001]
Precuneus	−.010 (.019)	−.006 (.019)	2.147	.032	−0.202	−.004	[−.007, .000]
Medial orbitofrontal sulcus	−.013 (.032)	−.007 (.029)	1.958	.051	−0.184	−.006	[−.011, .000]
Pars orbitalis	−.006 (.043)	.002 (.039)	1.912	.056	−0.179	−.007	[−.015, .000]
Insula	−.017 (.016)	−.014 (.015)	1.742	.082	−0.163	−.003	[−.005, .000]
Calcarine sulcus	.009 (.019)	.006 (.022)	−1.577	.116	0.148	.003	[−.001, .007]
Post−central gyrus	.003 (.029)	.007 (.026)	1.517	.130	−0.142	−.004	[−.009, .001]
Parahippocampal gyrus	−.023 (.021)	−.020 (.020)	1.505	.133	−0.141	−.003	[−.007, .001]
Inferior parietal lobule	−.011 (.030)	−.008 (.026)	1.475	.141	−0.138	−.004	[−.009, .001]
Paracentral lobule	−.006 (.024)	−.002 (.022)	1.391	.165	−0.131	−.003	[−.007, .001]

Continued

Table 4. (*Continued*)

Region of interest	Reliable change group	No-Change Group	*t*	*p*	Cohen's *d*	Mean difference	95% confidence interval for the mean difference
Superior parietal gyrus	−.003 (.032)	.000 (.030)	1.274	.203	−0.120	−.004	[−.009, .002]
Transverse temporal gyrus	−.014 (.021)	−.012 (.026)	−0.855	.393	0.080	.002	[−.002, .006]
Lateral occipital sulcus	−.006 (.028)	−.004 (.026)	0.813	.417	−0.076	−.002	[−.007, .003]
Rostral anterior cingulate cortex	−.012 (.019)	−.011 (.019)	0.358	.721	−0.034	−.001	[−.004, .003]
Lingual gyrus	−.003 (.015)	−.003 (.014)	−0.241	.809	0.023	.000	[−.002, .003]

convention for qualitatively labelling effect sizes (Cohen, 1988), these are considered 'small', even those for which the percent atrophy rate appears to be quite large.

Discussion

The goal of this project was to determine the criterion validity of the RCI when applied to composite cognitive measures of memory and executive functioning, using neuroimaging evidence of brain atrophy over a one-year follow-up period as the criterion standard. The current study built upon a previous pioneering study (Duff *et al.*, 2019) and employed a larger sample size, a longer retest period, matched samples and comprehensive longitudinal brain measurements. We were thus able to validate the RCI against actual brain volume change as opposed to baseline cross-sectional measures. In both the ADNI-Mem and the ADNI-EF studies, voxelwise whole brain analysis and ROI analysis revealed greater brain atrophy in the RC groups compared to the NC groups, which were matched on key confounding variables. Overall, the brain regions that demonstrated the greatest atrophy, evident in both studies, included the medial temporal cortex, the temporal lobe and some parts of the parietal lobe. The RC group in the ADNI-EF study demonstrated more extensive atrophy in the frontal and temporal lobes and posterior cingulate cortex compared to the RC group in the ADNI-Mem study. These findings demonstrate that the presence of reliable change on neuropsychological test scores can be valuable in making inferences about the underlying brain regions that are likely to have experienced atrophy over a given time interval.

Our results suggest that patterns of atrophy underlying reliable change on ADNI-Mem include the medial temporal cortex, the temporal lobe and some parts of the parietal lobe. In fact, the evidence showing the relationship between the medial temporal lobes and episodic memory functioning is abundant: the hippocampus, the parahippocampal gyrus and the entorhinal cortex are three brain regions that tend to be most highly correlated with episodic memory functioning. For example, a longitudinal study revealed that atrophy of the entorhinal cortex was predictive of worse memory performance five years later (Rodrigue & Raz, 2004). More broadly, there is abundant evidence in the literature showing that subcomponents of the limbic system, such as the fornix (Aggleton *et al.*, 2000), the mammillary bodies (Tsivilis *et al.*, 2008) and the cingulate gyrus (Burianova, McIntosh, & Grady, 2010; Maddock, Garrett, & Buonocore, 2001) are associated with new learning and retention. However, in the current study, when groups were defined on the basis of reliable change in ADNI-Mem scores, no group differences were detected in the hippocampus, a medial temporal lobe structure believed to be of primary importance for episodic memory ability. This unexpected finding is discussed in more detail below.

Results from the ADNI-EF study were different from, but overlapping with, patterns of atrophy in the ADNI-Mem study. Brain regions whose atrophy differed depending on the absence or presence of reliable change in ADNI-EF scores were the medial temporal cortex, portions of the lateral temporal lobe, portions of the frontal lobe and some regions of the parietal lobe. Compared to the ADNI-Mem results, reliable change in ADNI-EF scores was predictive of more atrophy in general, but especially in the temporal lobe, corpus callosum, medial orbitofrontal, thalamus and posterior cingulate cortex (Figure 2). Evidence of greater frontal lobe involvement is also seen in Table 4. The atrophy associated with reliable change in ADNI-EF is consistent with existing literature documenting the important role of the frontal lobes in promoting executive functions, both from lesion studies (Barceló & Knight, 2002; Eslinger & Grattan, 1993; van den

Broek, Bradshaw, & Szabadi, 1993) and from neuroimaging studies (Konishi *et al*., 1998; Mentzel *et al*., 1998; Volz *et al*., 1997). However, it is also worth mentioning that recent studies have focused more on neural connectivity integrating the frontal lobe to other regions of the brain in performing executive function tasks (Braun *et al*., 2015). Regions that connect with the frontal lobe for executive functioning include the parietal lobe, limbic system, basal ganglia, and other cortical and subcortical regions (for a comprehensive meta-analytic review, see Niendam *et al*., 2012). The fact that the current study also found concurrent atrophy in parietal lobe, temporal lobe and limbic system structures, in addition to the frontal lobe, supports the importance of neural connectivity in executive functions and Braun *et al*'s view, 2015 on frontal network integration.

The finding that the hippocampus did not show a meaningful difference in atrophy between the two groups in the ADNI-Mem and the ADNI-EF study was unanticipated. The results seem to indicate that both the RC (in ADNI-Mem: $M = .034$, $SD = .050$; in ADNI-EF: $M = .043$, $SD = .051$) and the NC groups (in ADNI-Mem: $M = .029$, $SD = .049$; in ADNI-EF: $M = .027$, $SD = .047$) showed *increased* hippocampal volumes at one-year follow-up compared to baseline. We believe that this finding is inaccurate. Because of varying degrees of hippocampal atrophy and the resulting morphological variability of the hippocampus in older populations, the B-spline deformation to template space is noisier in this region than in other brain regions. A primary contributor to this increased noise is the proximity of ventricles, which expand at a rate proportional to the atrophic changes in surrounding tissue; the resulting positive log-Jacobians of the ventricles mix with hippocampal tissue in the template space in cases of poor B-spline matches (Nestor *et al*., 2008). A way around this ROI analysis is to examine log-Jacobian means on carefully segmented hippocampal masks in native space. For the current study, these data were available only on a small number of our subjects. However, we were able to verify that hippocampal atrophy measured in this way was greater for change than no-change subjects in this small subset (data not shown). This result on a partial subset of our data is consistent with the Duff *et al*. (2019) study, which reported that reliable change on a battery of cognitive tests was associated with smaller hippocampal volumes (measured cross-sectionally). Still, further replication is warranted with regard to our unexpected findings regarding change in hippocampal volumes.

Our results make a unique contribution to the literature by providing criterion validation for the RCI as a clinical indicator of underlying brain volume changes. The current results are highly applicable to clinical situations where the underlying brain change is not known to the clinician, as is typical for many neuropsychological evaluations. In fact, many neuropsychologists perform longitudinal cognitive assessment for the purpose of making inferences about potential changes to the brain in the absence of serial MRI scans. The approach to documenting reliable change is based on the need to identify whether an observed change is larger than would be expected on the basis of measurement error; such a clinical decision is usually dichotomous. In this regard, the current study can contribute to clinical practice by validating reliable test score changes against parallel neuroanatomical outcomes. Thus, in the absence of patient neuroimaging data, these results may allow neuropsychologists to make inferences about expected patterns of brain atrophy when reliable decline is found on cognitive test scores.

Another contribution of the present research is that it demonstrates the clinical relevance of the RCI. The statistical procedures used to generate RCIs are based on the same approach used to perform null hypothesis significance testing. In essence, 'reliable' changes are changes that would be unlikely to occur simply as a result of measurement error. However, as with null hypothesis significance testing, these methods are purely

statistical; in and of themselves, they do not provide meaningful data about practical utility or clinical meaning (Millis, 2003). The current study moves beyond statistical significance and into the realm of clinical significance. Here, we demonstrate that a likely mechanism underlying reliable change in neuropsychological test scores is regional brain atrophy.

Finally, by our focus on categorical classifications of reliable change, we are validating not only the RCI formula against neuroanatomical outcomes, but the confidence threshold of this approach used to separate reliable change from no change (i.e., the one-tailed 90% confidence limit). Since all such thresholds necessarily include a level of arbitrary characterization, our work provides validation that in the current usage this thresholding corresponds to real differences in brain change over a one-year period. Nevertheless, evaluating brain associations to the continuous z-scores also generated by the simple regression-based approach to reliable change is another interesting and important topic that should be explored in future work.

The current research project has several limitations. We have already discussed the anomalous results for hippocampal atrophy, which resulted from limitations in the B-spline deformation of the template-space voxelwise approach. Among other limitations, a primary one is our demographic makeup. The majority of the participants in the current research were highly educated Caucasian Americans. This limitation is relatively common in dementia research in the United States, where racial and ethnic minorities are often underrepresented in ageing and dementia studies. What makes such limitations vexing is the accumulating evidence that race and ethnicity are important factors influencing the prevalence, aetiology and onset of dementia (Gavett *et al*., 2018; O'Bryant *et al*., 2013). For instance, African Americans are at a higher risk of developing dementia, almost two to four times higher than their Caucasian counterparts (Steenland, Goldstein, Levey, & Wharton, 2016). Despite the scarcity of racial and ethnical minorities in the current sample, the genetic matching procedure (Sekhon, 2011) ensured that racial and ethnical differences did not come into play as a confounding variable when comparing the RC and NC groups. Nevertheless, averaging participants' brain volume data can potentially obscure the influence of race and ethnicity for two obvious reasons. For one, by averaging all participants' data in aggregate, the disparity between racial and ethnic groups is obscured, meaning that variability due to race and ethnicity cannot be explored. For the other, it is questionable how much the current findings can be generalized to minority races and ethnicities when less than 10% of the participants in each study were non-Hispanic Whites.

The problem with averaging participants' brain volume data despite their disparity on key characteristics can also be discussed in terms of the diagnosis of dementia. Even though the genetic matching procedure ensured equivalence of diagnostic membership across groups, the representation of clinical disease severity within groups was disproportionate when comparing the participants in the ADNI-Mem study to those in the ADNI-EF study. In other words, the participants in the ADNI-EF study were more likely to be diagnosed with late MCI or AD than those in the ADNI-Mem study. This difference between the two study samples could explain why the patterns of atrophy appeared more extensive in the context of reliable change in ADNI-EF compared to reliable change in ADNI-Mem. The impact of this potential confound across studies is dampened, however, by the fact that dementia severity measures (e.g., MMSE, CDR) were very similar in both study samples despite differences in diagnostic frequencies. It should be noted that in ADNI, diagnosis is determined by clinical application of standard diagnostic criteria (Petersen *et al*., 2010). Levels of MCI (early and late) were differentiated by scores on the delayed Logical Memory subtest from the Wechsler Memory Scale-Revised; in ADNI-1, all

participants with MCI had late MCI (Aisen *et al.*, 2010). In addition, reliable change occurring in cognitively normal older adults or those with early MCI at baseline may have different clinical implications than reliable change in those diagnosed with baseline diagnoses of AD or late MCI, which is a potentially intriguing topic yet to be explored.

One further limitation relates to the fact that slightly less than 10% of the participants in our main analyses provided data for the preliminary analyses needed to derive test–retest reliability, sample means and sample standard deviations at baseline and follow-up. The data derived from these preliminary analyses were used to assign participants into the RC or NC groups. Therefore, for a subset of our sample, there may have been a small amount of criterion contamination that could have affected our results. However, because such a small proportion of our sample was affected, we do not believe that this alters the data or our interpretation in a meaningful way.

Despite the lack of experimental manipulation in this observational research, the use of the genetic matching procedure strengthens the ability to attribute group differences in brain atrophy to the cognitive manifestations of reliable change, rather than to other possible confounding variables (e.g., differences in age or baseline brain volume). Despite the fact that, in the ADNI-EF study, a match on age was not achieved, it should not represent a major limitation, as the difference between groups was small (1.56 years) and the NC group was slightly older.

The data used in the current research were obtained from ADNI; therefore, the results came primarily from an older adult sample enriched for likely AD pathology and not other dementia aetiologies. The fact that the sample consists of mostly older adults with AD or at risk for AD is both a weakness and a strength of the current project. It is a weakness because the findings cannot be generalized to other types of dementia, such as frontotemporal dementia. Yet, it is also a strength of the current project; because AD is the most prevalent cause of dementia (Brookmeyer, Johnson, Ziegler-Graham, & Arrighi, 2007), the current findings can be generalized to a large number of patients with known or suspected AD.

As the neuropsychological test data used here to identify reliable change came from research-based factor scores of memory and executive functioning, future research should extend these findings to apply the same construct validation approach to observed test scores obtained from standard neuropsychological instruments. For example, examining the criterion validity of RCIs derived from common clinical test scores, and perhaps comparing these to factor scores is an obvious first step. In fact, our group is currently pursuing this as a follow-up study; still, additional research is needed.

Conclusion

The current results demonstrate the criterion validity of the RCI for corresponding to meaningful patterns of brain change in older adults over the span of one year of follow-up. Reliable change on ADNI-Mem was associated with greater atrophy in the medial temporal cortex, temporal lobe and some regions of the parietal lobe compared to the NC group. Reliable change on ADNI-EF was associated with greater atrophy in the medial temporal cortex, posterior corpus callosum and cingulate cortex, temporal lobe, some regions of the parietal lobe and the frontal lobe compared to the NC group. The current results suggest that reliable change in these factor scores has criterion validity for mapping onto expected changes in the underlying brain structure. More research using this approach can have clinical value for making inferences about the possible brain changes underlying reliable test score changes.

Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; Bioclinica, Inc.; Biogen; Bristol Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; Euroimmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Conflicts of interest

All authors declare no conflict of interest.

Author contributions

Shayne S.-H. Lin, M.A. (Conceptualization; Formal analysis; Investigation; Methodology; Project administration; Writing – original draft; Writing – review & editing) Evan Fletcher (Data curation; Formal analysis; Software; Visualization; Writing – review & editing) Brandon E. Gavett (Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Resources; Software; Supervision; Writing – review & editing)

References

Aggleton, J. P., McMackin, D., Carpenter, K., Hornak, J., Kapur, N., Halpin, S., . . . Gaffan, D. (2000). Differential cognitive effects of colloid cysts in the third ventricle that spare or compromise the fornix. *Brain*, *123*, 800–815. https://doi.org/10.1093/brain/123.4.800

Aisen, P. S., Petersen, R. C., Donohue, M. C., Gamst, A., Raman, R., Thomas, R. G., . . . Weiner, M. W. (2010). Clinical core of the Alzheimer's disease neuroimaging initiative: Progress and plans. *Alzheimer's & Dementia*, *6*, 239–246. https://doi.org/10.1016/j.jalz.2010.03.006

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Bakkour, A., Morris, J. C., & Dickerson, B. C. (2009). The cortical signature of prodromal AD: Regional thinning predicts mild AD dementia. *Neurology*, *72*, 1048–1055. https://doi.org/10.1212/01.wnl.0000340981.97664.2f

Barceló, F., & Knight, R. T. (2002). Both random and perseverative errors underlie WCST deficits in prefrontal patients. *Neuropsychologia*, *40*, 349–356. https://doi.org/10.1016/S0028-3932(01)00110-5

Bowden, S. C., & Finch, S. (2017). When is a test reliable enough and why does it matter? In S. C. Bowden (Ed.), *Neuropsychological assessment in the age of evidence-based practice diagnostic and treatment evaluations* (pp. 95–119). New York, NY: Oxford University Press.

Braun, U., Schäfer, A., Walter, H., Erk, S., Romanczuk-Seiferth, N., Haddad, L., . . . Meyer-Lindenberg, A. (2015). Dynamic reconfiguration of frontal brain networks during executive cognition in humans. *Proceedings of the National Academy of Sciences*, *112*, 11678–11683. https://doi.org/10.1073/pnas.1422487112

Brookmeyer, R., Johnson, E., Ziegler-Graham, K., & Arrighi, H. M. (2007). Forecasting the global burden of Alzheimer's disease. *Alzheimer's & Dementia*, *3*, 186–191. https://doi.org/10.1016/j.jalz.2007.04.381

Burianova, H., McIntosh, A. R., & Grady, C. L. (2010). A common functional brain network for autobiographical, episodic, and semantic memory retrieval. *NeuroImage*, *49*(1), 865–874. https://doi.org/10.1016/j.neuroimage.2009.08.066

Chelune, G. J., Naugle, R. I., Lüders, H., Sedlak, J., & Awad, I. A. (1993). Individual change after epilepsy surgery: Practice effects and base-rate information. *Neuropsychology*, *7*(1), 41–52. https://doi.org/10.1037/0894-4105.7.1.41

Christensen, L., & Mendoza, J. L. , (1986). A method of assessing change in a single subject: An alteration of the RC index. *Behavior Therapy*, *17*(3), 305–308. http://dx.doi.org/10.1016/s0005-7894(86)80060-0

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd ed. Hillsdale, NJ: Lawrence-Erlbaum.

Crane, P. K., Carle, A., Gibbons, L. E., Insel, P., Mackin, R. S., Gross, A., . . . Weiner, M. (2012). Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Brain imaging and behavior*, *6*, 502–516. https://doi.org/10.1007/s11682-012-9186-z

Cummings, J. L., Tomiyasu, U., Read, S., & Benson, D. F. (1984). Amnesia with hippocampal lesions cardiopulmonary arrest. *Neurology*, *34*, 679. https://doi.org/10.1212/WNL.34.5.679

Danet, L., Barbeau, E. J., Eustache, P., Planton, M., Raposo, N., Sibon, I., . . . Pariente, J. (2015). Thalamic amnesia after infarct: the role of the mammillothalamic tract and mediodorsal nucleus. *Neurology*, *85*, 2107–2115. https://doi.org/10.1212/WNL.0000000000002226

Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, *95*, 932–945. https://doi.org/10.1162/REST_a_00318

Dickerson, B. C., Bakkour, A., Salat, D. H., Feczko, E., Pacheco, J., Greve, D. N., Grodstein, F., Wright, C. I., Blacker, D., Rosas, H. D., Sperling, R., Atri, A., Growdon, J. H., Hyman, B. T., Morris, J. C., Fischl, B., & Buckner, R. L. (2009). The cortical signature of Alzheimer's disease: Regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals. *Cerebral Cortex*, *19*, 497–510. https://doi.org/10.1093/cercor/bhn113

Duff, K., Suhrie, K. R., Dalley, B. C., Anderson, J. S., & Hoffman, J. M. (2019). External validation of change formulae in neuropsychology with neuroimaging biomarkers: A methodological recommendation and preliminary clinical data. *The Clinical Neuropsychologist*, *33*, 478–489. https://doi.org/10.1080/13854046.2018.1484518

Eslinger, P. J., & Grattan, L. M. (1993). Frontal lobe and frontal-striatal substrates for different forms of human cognitive flexibility. *Neuropsychologia*, *31*(1), 17–28. https://doi.org/10.1016/0028-3932(93)90077-D

Fletcher, E., Gavett, B., Harvey, D., Farias, S. T., Olichney, J., Beckett, L., . . . Mungas, D. (2018). Brain volume change and cognitive trajectories in aging. *Neuropsychology*, *32*, 436. https://doi.org/10.1037/neu0000447

Fletcher, E., Knaack, A., Singh, B., Lloyd, E., Wu, E., Carmichael, O., & DeCarli, C. (2013). Combining boundary-based methods with tensor-based morphometry in the measurement of longitudinal brain change. *IEEE Transactions on Medical Imaging*, *32*, 223–236. https://doi.org/10.1109/TMI.2012.2220153

Fletcher, E., Singh, B., Harvey, D., Carmichael, O., & DeCarli, C. (2012). Adaptive image segmentation for robust measurement of longitudinal brain tissue change. In Conference proceedings:... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference (Vol. 2012, p. 5319). NIH Public Access. https://doi.org/10.1109/EMBC.2012.6347195.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*, 189–198. https://doi.org/10.1016/0022-3956(75)90026-6

Gavett, B. E., Fletcher, E., Harvey, D., Farias, S. T., Olichney, J., Beckett, L., ... Mungas, D. (2018). Ethnoracial differences in brain structure change and cognitive change. *Neuropsychology*, *32*, 529. https://doi.org/10.1037/neu0000452

Gibbons, L. E., Carle, A. C., Mackin, R. S., Harvey, D., Mukherjee, S., Insel, P., ... Alzheimer's Disease Neuroimaging Initiative (2012). A composite score for executive functioning, validated in Alzheimer's Disease Neuroimaging Initiative (ADNI) participants with baseline mild cognitive impairment. *Brain Imaging and Behavior*, *6*, 517–527. https://doi.org/10.1007/s11682-012-9186-z

Gunning-Dixon, F. M., & Raz, N. (2003). Neuroanatomical correlates of selected executive functions in middle-aged and older adults: a prospective MRI study. *Neuropsychologia*, *41*, 1929–1941. https://doi.org/10.1016/S0028-3932(03)00129-5

Hamani, C., McAndrews, M. P., Cohn, M., Oh, M., Zumsteg, D., Shapiro, C. M., ... Lozano, A. M. (2008). Memory enhancement induced by hypothalamic/fornix deep brain stimulation. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, *63*(1), 119–123. https://doi.org/10.1002/ana.21295

Hua, X., Leow, A. D., Parikshak, N., Lee, S., Chiang, M. C., Toga, A. W., ... Alzheimer's Disease Neuroimaging Initiative (2008). Tensor-based morphometry as a neuroimaging biomarker for Alzheimer's disease: an MRI study of 676 AD, MCI, and normal subjects. *NeuroImage*, *43*, 458–469. https://doi.org/10.1016/j.neuroimage.2008.07.013

Jacobson, N. S., & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12. https://doi.org/10.1037/0022-006X.59.1.12

Kaplan, E. (1988). A process approach to neuropsychological assessment. In T. Boll & B. K. Bryant (Eds.), *Clinical neuropsychological and brain function: Research, measurement, and practice*, (127–167). Washington, DC: American Psychological Association. https://doi.org/10.1037/10063-004

Kerchner, G. A., Racine, C. A., Hale, S., Wilheim, R., Laluz, V., Miller, B. L., & Kramer, J. H. (2012). Cognitive processing speed in older adults: relationship with white matter integrity. *PLoS One*, *7* (11), e50425. https://doi.org/10.1371/journal.pone.0050425

Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., ... Keshavan, A. (2017). Mindboggling morphometry of human brains. *PLoS Computational Biology*, *13*, e1005350. https://doi.org/10.1371/journal.pcbi.1005350

Kochunov, P., Lancaster, J. L., Thompson, P., Woods, R., Mazziotta, J., Hardies, J., & Fox, P. (2001). Regional spatial normalization: toward an optimal target. *Journal of computer assisted tomography*, *25*, 805–816. https://doi.org/10.1097/00004728-200109000-00023

Konishi, S., Nakajima, K., Uchida, I., Kameyama, M., Nakahara, K., Sekihara, K., & Miyashita, Y. (1998). Transient activation of inferior prefrontal cortex during cognitive set shifting. *Nature Neuroscience*, *1*(1), 80–84. https://doi.org/10.1038/283

Maddock, R. J., Garrett, A. S., & Buonocore, M. H. (2001). Remembering familiar people: the posterior cingulate cortex and autobiographical memory retrieval. *Neuroscience*, *104*, 667–676. https://doi.org/10.1016/S0306-4522(01)00108-7

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, Jr, C. R., Kawas, C. H., ... Mohs, R. C. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for

Alzheimer's disease. *Alzheimer's & Dementia*, *7*, 263–269. https://doi.org/10.1016/j.jalz.2011.03.005

McSweeny, A. J., Naugle, R. I., Chelune, G. J., & Luders, H. (1993). T scores for change: An illustration of a regression approach to depicting change in clinical neuropsychology. *The Clinical Neuropsychologist*, *7*, 300–312. https://doi.org/10.1080/13854049308401901

Mentzel, H. J., Gaser, C., Volz, H. P., Rzanny, R., Häger, F., Sauer, H., & Kaiser, W. A. (1998). Cognitive stimulation with the Wisconsin Card Sorting Test: functional MR imaging at 1.5 T. *Radiology*, *207*, 399–404. https://doi.org/10.1148/radiology.207.2.9577487

Millis, S. R. (2003). Emerging standards in statistical practice: implications for clinical trials in rehabilitation medicine. *American Journal of Physical Medicine & Rehabilitation*, *82*, S32–S37. https://doi.org/10.1097/01.PHM.0000087007.19214.32

Nestor, S. M., Rupsingh, R., Borrie, M., Smith, M., Accomazzi, V., Wells, J. L., . . .Alzheimer's Disease Neuroimaging Initiative (2008). Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database. *Brain*, *131*, 2443–2454. https://doi.org/10.1093/brain/awn146

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, *15*(1), 1–25. https://doi.org/10.1002/hbm.1058

Niendam, T. A., Laird, A. R., Ray, K. L., Dean, Y. M., Glahn, D. C., & Carter, C. S. (2012). Meta-analytic evidence for a superordinate cognitive control network subserving diverse executive functions. *Cognitive, Affective, & Behavioral Neuroscience*, *12*(2), 241–268. https://doi.org/10.3758/s13415-011-0083-5

O'Bryant, S. E., Johnson, L., Balldin, V., Edwards, M., Barber, R., Williams, B., . . . Hall, J. (2013). Characterization of Mexican Americans with mild cognitive impairment and Alzheimer's disease. *Journal of Alzheimer's Disease*, *33*(2), 373–379. https://doi.org/10.3233/JAD-2012-121420

Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., Harvey, D. J., . . . Weiner, M. W. (2010). Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization. *Neurology*, *74*(3), 201–209. https://doi.org/10.1212/WNL.0b013e3181cb3e25

Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., & Ferri, C. P. (2013). The global prevalence of dementia: a systematic review and metaanalysis. *Alzheimer's & Dementia*, *9*(1), 63–75. https://doi.org/10.1016/j.jalz.2012.11.007

R Core Team (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Reitan, R. M., & Wolfson, D. (1993). *The Halstead-Reitan Neuropsychological Battery: Theory and clinical interpretation*, 2nd ed., Tuscon, AZ: Neuropsychology Press.

Rey, A. (1941). L'examen psychologique dans les cas d'encéphalopathie traumatique (Les problems). *Archives de Psychologie*, *28*, 286–340.

Rodrigue, K. M., & Raz, N. (2004). Shrinkage of the entorhinal cortex over five years predicts memory performance in healthy adults. *Journal of Neuroscience*, *24*, 956–963. https://doi.org/10.1523/JNEUROSCI.4166-03.2004

Rosen, W. G., Mohs, R. C., & Davis, K. L. (1984). A new rating scale for Alzheimer's disease. *The American Journal of Psychiatry*, *141*, 1356–1364. https://doi.org/10.1176/ajp.141.11.1356

Rueckert, D., Aljabar, P., Heckemann, R. A., Hajnal, J. V., & Hammers, A. (2006). Diffeomorphic registration using B-splines. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 702–9). Berlin, Heidelberg: Springer. https://doi.org/10.1007/11866763_86

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *Journal of Statistical Software*, *42*, 1–52. https://doi.org/10.18637/jss.v042.i07

Steenland, K., Goldstein, F. C., Levey, A., & Wharton, W. (2016). A meta-analysis of Alzheimer's disease incidence and prevalence comparing African-Americans and caucasians. *Journal of Alzheimer's Disease*, *50*(1), 71–76. https://doi.org/10.3233/JAD-150778

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A Compendium of Neuropsychological Tests*, 3rd ed., New York, NY: Oxford University Press.

Thurstone, L. L. (1938). *Primary mental abilities. Psychometric monographs*. Chicago, IL: University of Chicago Press.

Tsivilis, D., Vann, S. D., Denby, C., Roberts, N., Mayes, A. R., Montaldi, D., & Aggleton, J. P. (2008). A disproportionate role for the fornix and mammillary bodies in recall versus recognition memory. *Nature Neuroscience*, *11*, 834. https://doi.org/10.1038/nn.2149

van den Broek, M. D., Bradshaw, C. M., & Szabadi, E. (1993). Utility of the modified Wisconsin Card Sorting Test in neuropsychological assessment. *British Journal of Clinical Psychology*, *32*, 333–343. https://doi.org/10.1111/j.2044-8260.1993.tb01064.x

Volz, H. P., Gaser, C., Häger, F., Rzanny, R., Mentzel, H. J., Kreitschmann-Andermahr, I., . . . Sauer, H. (1997). Brain activation during cognitive stimulation with the Wisconsin Card Sorting Test—a functional MRI study on healthy volunteers and schizophrenics. *Psychiatry Research: Neuroimaging*, *75*, 145–157. https://doi.org/10.1016/S0925-4927(97)00053-X

Wechsler, D. (1981). *Manual for the wechsler adult intelligence scale-revised (WAIS-R)*. New York, NY: The Psychological Corporation.

Wechsler, D. (1987). *Wechsler memory scale — Revised*. New York, NY: Psychological Corporation.

Received 23 March 2020; revised version received 10 September 2020