NeuroImage 84 (2014) 443–452

Contents lists available at ScienceDirect

NeuroImage

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

Reliability measures of functional magnetic resonance imaging in a longitudinal evaluation of mild cognitive impairment

Theodore P. Zanto [a,*,1], Judy Pa [a,1], Adam Gazzaley [a,b]

[a] Department of Neurology, University of California San Francisco, San Francisco, CA 94158, USA
[b] Departments of Physiology and Psychiatry, University of California San Francisco, San Francisco, CA 94158, USA

ARTICLE INFO

Article history:
Accepted 29 August 2013
Available online 7 September 2013

Keywords:
Functional magnetic resonance imaging
Mild cognitive impairment
Intraclass correlation
Test–retest reliability
Working memory
Selective attention

ABSTRACT

As the aging population grows, it has become increasingly important to carefully characterize amnestic mild cognitive impairment (aMCI), a preclinical stage of Alzheimer's disease (AD). Functional magnetic resonance imaging (fMRI) is a valuable tool for monitoring disease progression in selectively vulnerable brain regions associated with AD neuropathology. However, the reliability of fMRI data in longitudinal studies of older adults with aMCI is largely unexplored. To address this, aMCI participants completed two visual working tasks, a Delayed-Recognition task and a One-Back task, on three separate scanning sessions over a three-month period. Test–retest reliability of the fMRI blood oxygen level dependent (BOLD) activity was assessed using an intraclass correlation (ICC) analysis approach. Results indicated that brain regions engaged during the task displayed greater reliability across sessions compared to regions that were not utilized by the task. During task-engagement, differential reliability scores were observed across the brain such that the frontal lobe, medial temporal lobe, and subcortical structures exhibited fair to moderate reliability (ICC = 0.3–0.6), while temporal, parietal, and occipital regions exhibited moderate to good reliability (ICC = 0.4–0.7). Additionally, reliability across brain regions was more stable when three fMRI sessions were used in the ICC calculation relative to two fMRI sessions. In conclusion, the fMRI BOLD signal is reliable across scanning sessions in this population and thus a useful tool for tracking longitudinal change in observational and interventional studies in aMCI.

Introduction

Over the next twenty years, the number of individuals 65 years and older is expected to double in the United States (Census-Bureau, 2008). Along with this growth in the size of the older population, the number of adults suffering from dementia will also increase. Thus, there is an important need to identify effective markers of disease progression in neurodegenerative disorders, such as Alzheimer's disease. Amnestic mild cognitive impairment is a preclinical stage of Alzheimer's disease (Petersen, 1995, 2000) and provides a unique window into the earliest disease-related changes. Longitudinal studies using functional magnetic resonance imaging (fMRI), a non-invasive, in vivo measure of brain function, have begun elucidating early brain changes in aMCI, but few studies have reported the test–retest reliability of the fMRI signal in this vulnerable population over time.

Despite the importance of evaluating the reliability of fMRI data in vulnerable populations, such as aMCI, this has only been assessed in a few clinical populations, such as Alzheimer's disease (Atri et al., 2011), stroke (Chen and Small, 2007; Eaton et al., 2008; Kimberley et al., 2008), schizophrenia (Manoach et al., 2001), focal epilepsy (Fernandez et al., 2003), and chronic nonfluent aphasia (Kurland et al., 2004). To our knowledge, only one study to date has assessed reliability in aMCI patients (Clement and Belleville, 2009), and this study, like most, utilized two fMRI sessions. It is unclear whether two fMRI sessions are sufficient to properly estimate reliability, especially in populations such as aMCI and Alzheimer's disease where neural degradation over time is a hallmark of the disease. Moreover, previous research of fMRI data reliability from aMCI and Alzheimer's populations focused on a select few brain regions of interest. Therefore, it is unclear whether neural activity in aMCI may be uniformly reliable across the brain or has region-specific differences in reliability.

Because little is known regarding the reliability of fMRI data in the aMCI population, this study aimed to: 1) systematically evaluate the spatial distribution of fMRI data reliability across the brain in aMCI, and 2) explore the stability of test–retest measurements as a factor of the number of fMRI sessions. To address this, test–retest reliability of fMRI blood oxygen level dependent (BOLD) activity was assessed in aMCI during three separate fMRI sessions while participants were engaged in two different visual working memory tasks (a Delayed Recognition and a One-Back task). The utility of using a Delayed Recognition and a One-Back task has been well documented, as they require the use of working memory processes that are affected in aMCI populations (e.g., Gomez-Tortosa et al., 2012; Missonnier et al., 2005, 2006).

[*] Corresponding author at: 675 Nelson Rising Ln., Box 0444, San Francisco, CA 94158, USA.
E-mail address: theodore.zanto@ucsf.edu (T.P. Zanto).
[1] Both authors contributed equally to this work.

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.08.063

Test–retest reliability was assessed via the intraclass correlation coefficient (ICC) (Shrout and Fleiss, 1979). ICC values were calculated across the brain from each of the two tasks. We hypothesized that the frontal lobe, medial temporal lobe, and subcortical structures would exhibit lower reliability measures, given the known susceptibility to atrophy and functional decline in aging and aMCI, when compared to less affected structures (Cherubini et al., 2010; De Vogelaere et al., 2012; Ferreira et al., 2011; Yang et al., 2012). Despite this relative difference, we hypothesized that these regions would still exhibit moderate or better levels of reliability across fMRI sessions. Additionally, we hypothesized that measurements of reliability would be more consistent when incorporating three, compared to two, fMRI sessions in the ICC calculation.

Methods

Participants

Thirteen older adults with mild memory deficits (age 63.8 ± 7.4 years; range 54–81 years of age; 5 females; Table 1) gave written informed consent to participate, which was approved by the University of California, San Francisco Committee for Human Research. Participants were recruited from the University of California, San Francisco Memory and Aging Center or through community screening. Note that all data analyzed in this study were obtained from participants that were a placebo control group in a larger study recently published (Pa et al., 2013). They were screened and diagnosed after an extensive neurological and neuropsychological evaluation. The one-hour neuropsychological screening battery assessed multiple domains of cognition, including memory, executive function, language, and visuospatial skills. Screening for depression was done using the self-reported 30-item Geriatric Depression Scale (GDS). Diagnosis of mild memory impairment was determined by consensus involving the neurologist and neuropsychology specialist. Participants had to endorse significant memory decline over the past year and demonstrate objective memory impairment (≥1SD below age- and education-matched normative values) on verbal or visual memory testing. Participants were excluded if they met criteria for dementia (DSM-IV), a history of a neurological disorder, current psychiatric illness or depression, head trauma with loss of consciousness greater than 10 min, severe sensory deficits, substance abuse, or were taking medications that affect cognition, such as donepezil. Participants were monetarily compensated for their participation and were offered an optional 3-month supply of donepezil after study completion (upon providing a physician's written prescription).

Table 1
Demographic and neuropsychological test performance at baseline. * indicates significant within-group differences from age- and education-matched normative values at $p < 0.05$. As expected based on diagnostic criteria, the memory scores were significantly lower than normative values. Values are presented as the mean and standard deviation in parentheses.

	Subject demographics	aMCI
	N (M/F)	13 (7/6)
	Age (years)	69.2 (8.2)
	Handedness (left/right)	2/11
	Education (years)	16.3 (2.0)
	Geriatric Depression Scale (GDS)	4.5 (3.9)
	Neuropsychological screening tests	
Global	MMSE (max. 30)	28.4 (1.9)
Memory	CVLT long delay free recall (max. 16)	5.4 (3.5)*
	Delayed Benson figure recall (max. 17)	7.9 (3.7)*
	Logical memory immediate (max. 25)	9.4 (3.8)*
	Logical memory delayed (max. 25)	6.6 (3.6)*
	Digit span backward	5.3 (1.3)
Attention/processing speed	Digit span forward	6.7 (1.1)
	WAIS-III digit symbol (90 s)	42.5 (8.7)
	Number sequencing (max. 150 s)	36.4 (12.1)
Executive function	Modified trailmaking Test B (max. 300 s)	83.4 (36.5)
	Stroop interference (number correct in 60 s)	41.3 (10.2)
	Verbal fluency (D words in 60 s)	14.9 (4.5)
Visuospatial	Copy of Benson figure (max. 17)	15.3 (1.2)

Participants completed a baseline fMRI session (Visit 1), a 1-month post-baseline fMRI session (Visit 2), and a 3-month post-baseline fMRI session (Visit 3). A total of 3 fMRI sessions were completed, and placebo pills were prescribed for the entire 3-month period after the baseline fMRI session.

Neuropsychological testing

Participants were administered a comprehensive screening battery of neuropsychological tests assessing memory, executive function, and visuospatial skill (summarized in Table 1). Tests of memory included the 20-minute delayed recall on California Verbal Learning Test (Delis et al., 2000), modified Logical memory 15-minute delay (Wechsler, 1987), and 10-minute recall of the Benson figure (Possin et al., 2011). The tests of executive function/attention included modified trail-making Test B (time to complete; Tombaugh, 2004), design fluency (number of unique designs in 60 s; Delis et al., 2001), modified stroop interference (number correct in 60 s; Stroop, 1935), letter fluency (D words in 60 s; Kramer et al., 2003), backward digit span (longest length; WAIS-R, Wechsler, 1981) and digit symbol (number correct in 60 s; WAIS-R, Wechsler, 1981). Tests of visuospatial function included constructional copy of the Benson figure.

Experimental design

For each of the three fMRI sessions, participants performed the same two visual working memory tasks (Fig. 1). Participants viewed the stimulus presentation monitor through a mirror located in front of their eyes. Stimuli were presented using E-Prime software (Psychology Software Tools, Sharpsburg, PA). Grayscale images of faces and natural scenes were used as stimuli. All cue images were novel throughout the fMRI experiment. Stimuli were 225 pixels wide by 300 pixels tall, and subtended approximately 5 by 6° of visual angle. Both male and female faces with neutral expressions were used, although the sex of the face stimuli used within each trial was held constant. The face stimuli were blurred along the contours of the faces, so that only the faces themselves were visible.

Delayed Recognition task

During the Delayed Recognition task (Fig. 1A), participants were presented two task conditions: Remember Faces/Ignore Scenes, and Remember Scenes/Ignore Faces. During each condition, participants were instructed to hold the relevant information in memory over a delay period and press a button indicating if the probe stimulus matched one of the two items previously presented (forced choice yes/no response). Probe stimuli matched on 50% of the trials. Conditions were presented in a counterbalanced order. A Passively View condition was also presented, but was not included in the current analysis. Data were acquired during 6 blocks lasting 4.5 min each, with each block containing 10 trials of one task condition. At the beginning of each block, instructions were presented to either Remember Faces (and ignore scenes) or Remember Scenes (and ignore faces). All face/scene stimuli were presented for 800 ms followed by a 200 ms blank screen with a central fixation cross.

One-Back task

Prior to the Delayed Recognition task, participants were presented a One-Back task for face and scene stimuli (Fig. 1B). Participants were instructed to press a button whenever the same stimulus (face or scene) was presented twice in a row. Face and scene stimuli were presented in alternating 16 s runs separated by an 8 s rest period for a total of 5 runs for each stimulus type. Each stimulus was presented for 400 ms followed by a 400 ms blank screen with a central fixation cross. This task was presented in one block lasting 4 min.

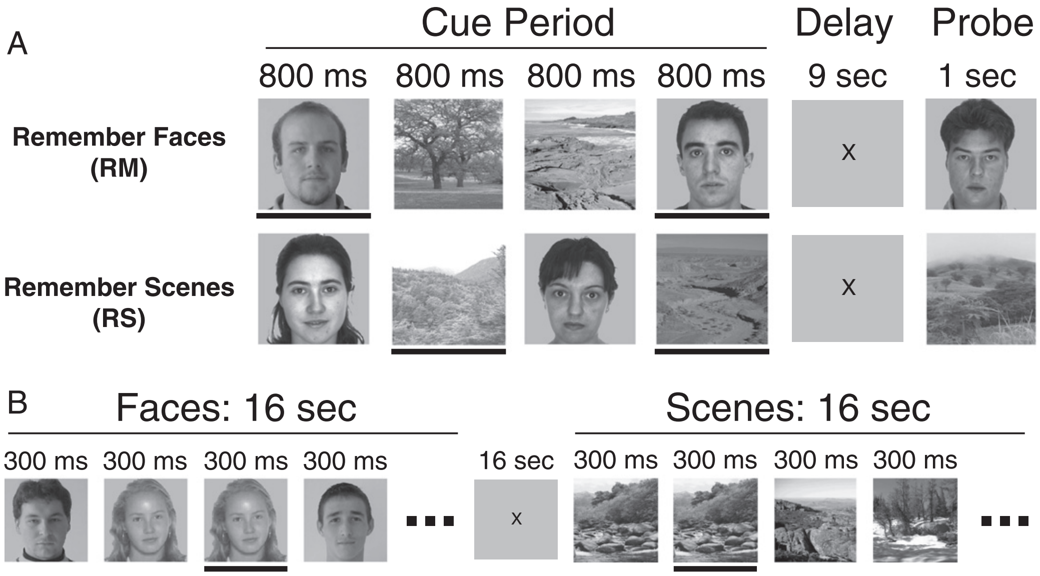

Fig. 1. Experimental tasks assessing working memory for face and scene stimuli. (A) Selective attention Delayed-Recognition task. A black bar below the stimulus indicates it was to be remembered and was not present during the task (display purposes only). (B) One-Back task. A black bar below the stimulus indicates it matched the previous stimulus and a button press was required. Again, this bar was not present during the task (display purposes only).

Data acquisition

All fMRI data was collected on a Siemens 3T MAGNETOM Trio with stimuli presented on an LCD monitor positioned behind the head of the participants and viewed using a mirror rigidly attached to a 12-channel head-coil. Echo planar imaging data was acquired (FA = 77^0, TE = 28 ms, TR = 2 s) with 29 interleaved axial slices and a 1.8 × 1.8 × 3 mm voxel size (FOV = 23 cm; 128 × 128 matrix). All pre-preprocessing of the data was conducted in SPM5 (Wellcome Department of Imaging Neuroscience, London, England). Raw blood oxygen level dependent (BOLD) data was corrected offline for slice-timing acquisition and motion-artifacts. A 5 mm isotropic Gaussian smoothing kernel was applied prior to modeling the data. To aid in anatomical localizations of BOLD activity, high-resolution T1-MPRAGE images were acquired (1 × 1 × 1 mm voxel size; FOV = 160 × 240 × 256 mm, TR = 2300 ms, TE = 3 ms, FA = 9^0). Each dataset was examined for excessive subject motion, greater than 2 mm (the size of a voxel), and no datasets were excluded.

Separate regressors were modeled for each stage of the trial (encode, delay, and probe) and convolved with a canonical Gaussian hemodynamic response function (SPM5, Wellcome Department of Imaging Neuroscience, London, England). The single instruction period at the start of each block was removed from the analysis. In addition, three translational (X, Y, Z) and three rotational (pitch, roll, yaw) motion parameters were included in the GLM to account for motion-related artifacts. The resulting regression vector yielded scalar beta weights corresponding to the relative changes in signal strength associated with a particular trial stage. Correct and incorrect trials were modeled separately and only correct trials were subjected to further analysis. Whole-brain BOLD activity maps were corrected for multiple comparisons by thresholding p-values with a cluster extent (i.e., 35 contiguous voxels exceeding $p < 0.01$) determined by a Monte Carlo simulation, yielding a corrected p-value of 0.05. Anatomical regions were identified as the most probable region within the Harvard–Oxford probabilistic atlas (Desikan et al., 2006). All analyses focused on the working memory encoding stage of the task.

Regions of interest

Three regions of interest (ROIs) were assessed for reliability: the fusiform face area (FFA) due to its responsivity to face stimuli (Kanwisher et al., 1997), the parahippocampal place area (PPA) for its sensitivity to scene stimuli (Epstein and Kanwisher, 1998), and the hippocampus because of its role in memory processes (reviewed in Squire et al., 2004) and selective vulnerability in aMCI patients (reviewed in Dickerson and Sperling, 2008). Each FFA and PPA ROI was defined as the cluster of 35 contiguous voxels within each anatomical region with the highest t-value on a face-scene for FFA and scene-face for PPA contrast from the One-Back task during the first fMRI session. All FFA and PPA ROIs were defined within each participant's native brain space. Additionally, group-level hippocampal ROIs were identified as the largest cluster of activity from each task within the hippocampus as defined by the Harvard–Oxford probabilistic atlas. Hippocampal activity was not elicited during the One-Back task and as such, ROI analyses were restricted to data from the Delayed Recognition task.

Intraclass correlation coefficient

Test–retest reliability was assessed via the intraclass correlation coefficient (ICC)(Shrout and Fleiss, 1979). The ICC model used was ICC(2,1), as defined by Shrout and Fleiss (1979), in which both the fMRI and participants are treated as random effects to assess reliability at a single point in time. This form of ICC utilizes a two-way ANOVA to estimate the correlation of BOLD activity between sessions. Significant BOLD activity was identified from the first fMRI visit and ICC values were calculated on those clusters of activity. Thus, ICC values reported are from regions that are task-based. Test–retest reliability was characterized as excellent (ICC > 0.8), good (ICC 0.6–0.79), moderate (ICC 0.4–0.59), fair (ICC 0.2–0.39), or poor (ICC < 0.2) (Guo et al., 2012). To assess the stability of ICC values based on a differential number of fMRI sessions, ICC was calculated from the first two sessions and compared to the ICC values from all three fMRI sessions. The first two sessions were selected, as it would be most similar to previous studies that utilize two fMRI sessions in calculating ICC.

Results

Delayed Recognition task

Behavior

Neuroimaging studies in clinical populations often aim to assess neural changes that are associated with declines in cognitive performance.

Therefore, it is important to address reliability of performance measures in the study population. Fig. 2 summarizes ICC results for both accuracy and response time when utilizing two and three experimental sessions. When incorporating data from three experimental sessions, results showed moderate reliability for accuracy (Fig. 2A) and good reliability for response times (Fig. 2B). Interestingly, when utilizing only the first two experimental sessions to estimate ICC, reliability of accuracy was different based on the specific Delayed Recognition task condition (Fig. 2C) whereas reliability of response times remained unchanged across conditions (Fig. 2D). Overall, these results indicate that response times are a more reliable measure of working memory performance across sessions than accuracy. Moreover, the 95% confidence intervals were smaller (upper–lower bound) when using three, compared to two, fMRI sessions to calculate ICC. This indicates that the additional fMRI session resulted in a better estimate of reliability.

Regions of interest

Activity in three regions of interest (ROIs) were assessed for test–retest reliability during the working memory encoding stage of each condition of the Delayed Recognition task across the three fMRI sessions: the fusiform face area (FFA; center of mass coordinates for left [x = −42, y = −58, z = −20], and right [x = 38, y = −52, z = −22] in MNI space), parahippocampal place area (PPA; left [x = −28, y = −44, z = −18], right [x = 24, y = −42, z = −18]), and the hippocampus (left [x = −22, y = −16, z = −16], right [x = 20, y = −16, z = −14]). Lateralized effects were not expected, nor were they observed, and so ICC values from the left and right ROIs were averaged together. BOLD activity during the Delayed Recognition task in the FFA and PPA exhibited attentional modulation such that attended stimuli elicited higher BOLD activity than ignored stimuli (Fig. 3). However, this relationship was not observed in the hippocampus, most likely because it is not known to be selective for specific types of visual stimuli. Both FFA and PPA exhibited good to excellent reliability measures (Fig. 4A, ICC 0.64–0.81) whereas the hippocampus elicits fair to moderate reliability (Fig. 4A, ICC 0.37–0.49). Given that BOLD activity in the FFA is less responsive to scene stimuli and activity in the PPA is less responsive to face stimuli, it could be surmised that reliability may decrease in neural regions that are task-irrelevant (i.e., FFA Remember Faces >> Remember Scenes; PPA Remember Scenes >> Remember Faces). This could be particularly true in older adults who typically exhibit a deficit in suppressing activity to irrelevant information (Gazzaley et al., 2005, 2008). However, both stimulus selective ROIs also exhibited good reliability measures when the stimulus was irrelevant (i.e., FFA for Remember Scenes, PPA for Remember Faces). Together, these results indicate that regardless of the stimulus relevance, stimulus selective ROIs, such as the FFA and PPA, yield more reliable BOLD activity than the hippocampus during working memory encoding.

Although similar results were observed when only using data from the first two fMRI sessions (Fig. 4B), reliability measures in the FFA and PPA were more contingent on the Delayed Recognition task condition such that the Remember Faces condition yielded higher ICC values compared to the Remember Scenes condition. ICC values in the hippocampus displayed similar differences between conditions, but a general increase in reliability when using data from three, compared to two, fMRI sessions. While the ICC to both accuracy and ROIs from two fMRI sessions showed a dependency on the Delayed Recognition task condition, they each showed different response profiles. Specifically, during the Remember Faces condition, ICC for accuracy increased with an extra fMRI session, whereas ICC for each ROI remained relatively unchanged. For the Remember Scenes condition, ICC for accuracy decreased with an extra fMRI session, while the ICC for each ROI increased. It was not expected that ICC values should differ between conditions and that by adding an extra scan in the ICC estimate results in more similar ICC values across tasks. Thus, the use of increased data in the ICC calculation may reflect a more stable measure of reliability. In support of this, the 95% confidence intervals were smaller (upper–lower bound) for each ROI and condition when three, compared to two, fMRI sessions were used. This corroborates the behavioral results to suggest that by incorporating a third fMRI session, a better estimate of reliability may be obtained.

Whole brain analysis

BOLD activity during working memory encoding was observed in fronto-parietal regions subserving executive control and working memory in addition to activity in occipital, temporal and subcortical regions during both the Remember Faces (Fig. 5, Table 2) and Remember Scenes

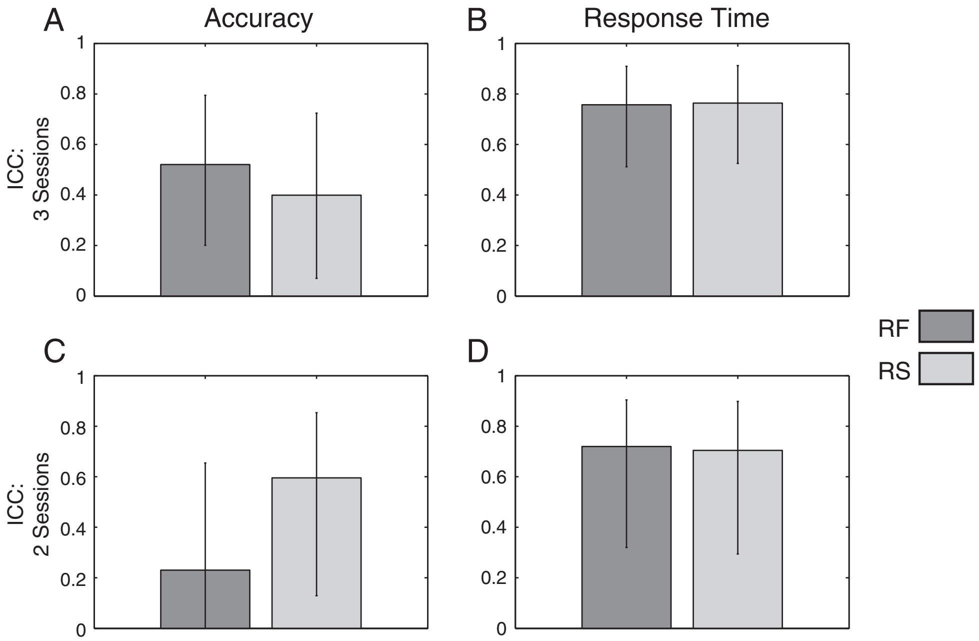

Fig. 2. Behavioral results. ICC values from A. accuracy and B. response times with 3 experimental sessions. ICC values from the first two experimental sessions for C. accuracy and D. response times. Error bars represent 95% confidence interval.

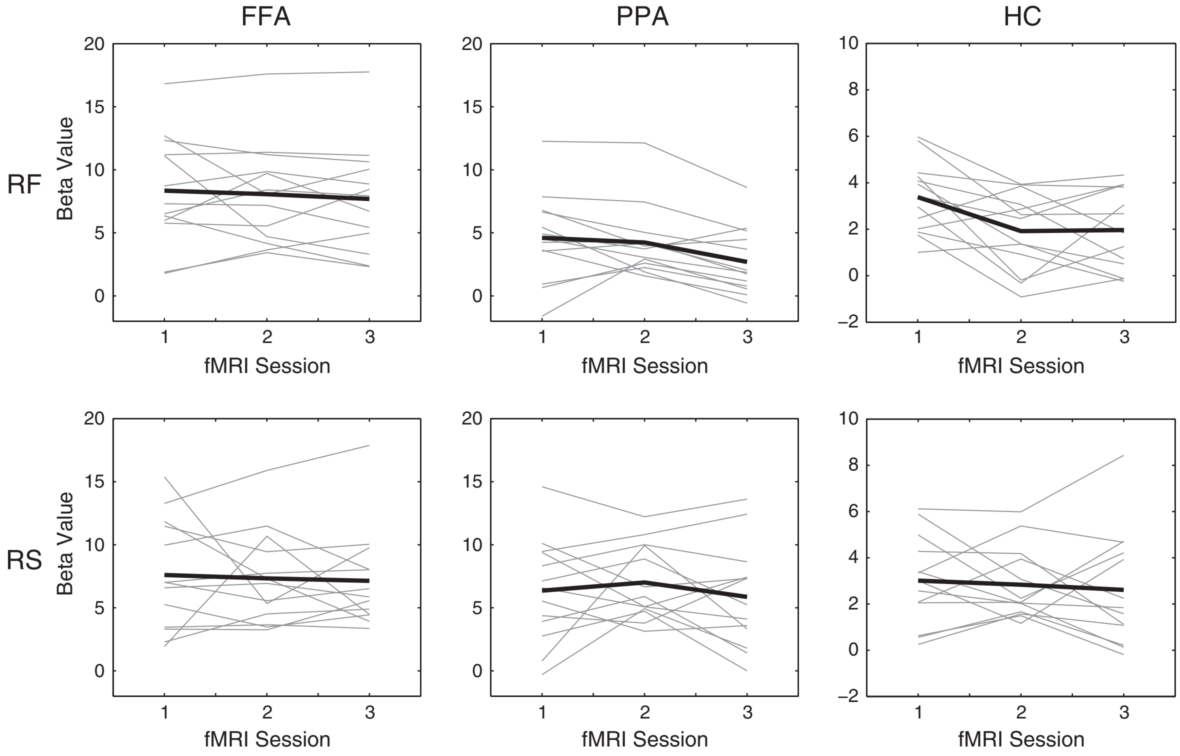

Fig. 3. BOLD activity in ROIs during the delayed recognition task. Light gray lines depict individual subject data. Dark black lines represent mean values.

(Fig. 5, Supplementary Table 1) conditions, and this activity was assessed for test–retest reliability via ICC. Each table lists regions with more than 100 voxels showing significant univariate BOLD activity during the task, their mean ICC value, the range of the 95% confidence intervals (CI), the change in CI range with three versus two fMRI sessions, how many of those voxels have a moderate ICC or better (i.e., >0.4), the percentage of univariate BOLD activity with a moderate ICC or better, and the percentage change between ICC calculated with three versus two fMRI sessions (calculated as the percentage of voxels where ICC > 0.4 when ICC is estimated from two fMRI sessions subtracted from the percentage of voxels where ICC > 0.4 when ICC is estimated from three fMRI sessions). Given the nature of the visual Delayed-Recognition task, it is not surprising that both the Remember Faces and Remember Scenes conditions yielded the greatest amount of univariate BOLD activity in the occipital lobe, followed by the frontal lobe, then temporal lobe, and the parietal lobe exhibited the least amount of BOLD activity (see Fig. 5 and the left column of Table 2 and Supplementary Table 1). Importantly, BOLD activity in the occipital lobe exhibited the greatest amount of moderate-to-excellent reliability (Remember Faces: 80%, Remember Scenes: 82%). Interestingly, the frontal lobe displayed the lowest amount of moderate-to-excellent reliability of all cortical lobes (Remember Faces: 51%, Remember Scenes: 60%), although the temporal lobe was only slightly better. Moreover, BOLD activity in subcortical structures exhibited the least amount of moderate-to-excellent reliability (Remember Faces: 26%, Remember Scenes: 36%). These results suggest that regions particularly susceptible to atrophy and functional decline in the aMCI population (i.e., frontal lobe, medial temporal lobe, and subcortical structures) (Cherubini et al., 2010; De Vogelaere et al., 2012; Ferreira et al., 2011; Yang et al., 2012) display less reliability for univariate BOLD activity relative to regions believed to be less affected in aMCI. In support of this, the range of the 95% confidence intervals was smallest within the occipital and parietal lobes, and greatest within the frontal, temporal and subcortical structures. Thus, as hypothesized, neural regions that are susceptible to decline in aMCI elicit less reliable BOLD activity.

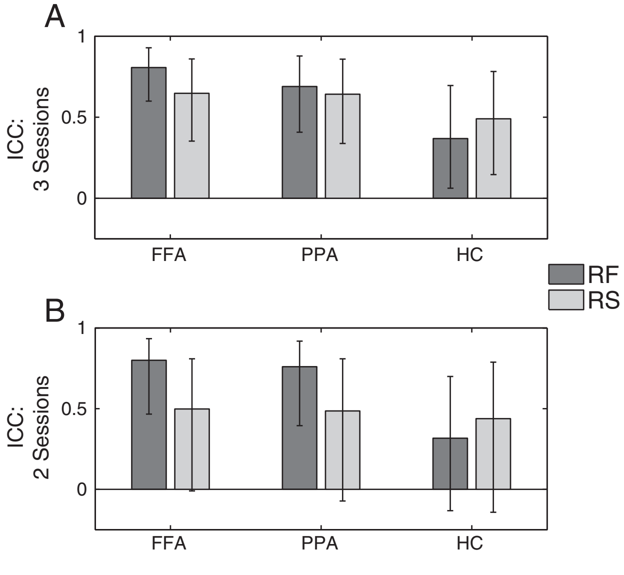

Fig. 4. Regions of interest for the fusiform face area (FFA), parahippocampal place area (PPA) and the hippocampus (HC) during the delayed recognition task for A. three fMRI sessions and B. two fMRI sessions. Error bars represent 95% confidence interval.

When using only the first two fMRI sessions to calculate ICC, there is a general decrease in the amount of BOLD activity that exhibits moderate-to-excellent reliability (Table 2 and Supplementary Table 1). Moreover, the 95% confidence intervals were observed to increase when using only two, compared to three, fMRI sessions in the ICC estimate, suggesting additional fMRI sessions result in better reliability estimates. However, the main findings remain such that the occipital lobe displays the highest amount of BOLD activity with moderate-to-excellent reliability whereas the frontal lobe (especially the anterior cingulate gyrus), medial temporal lobe (particularly the parahippocampal gyrus), and subcortical structures

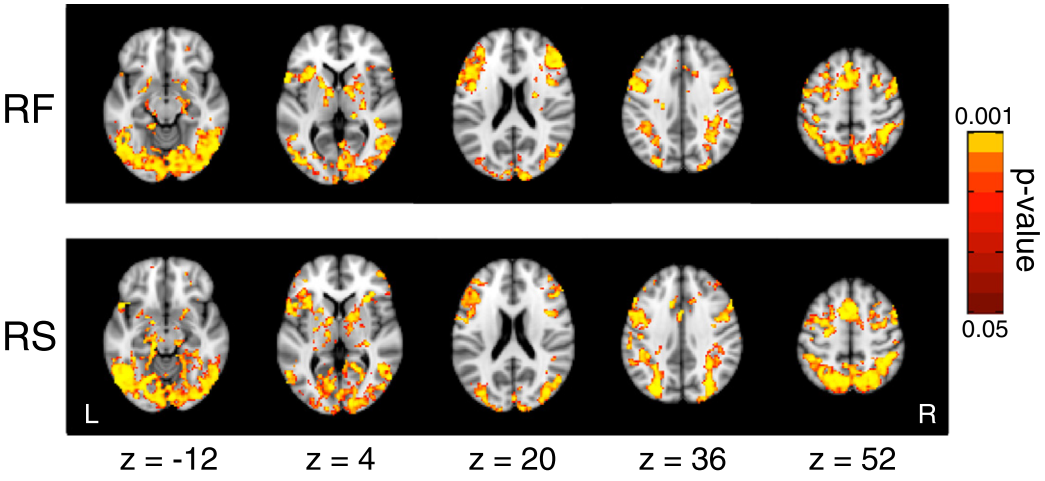

Fig. 5. Significant BOLD activity across the brain for each condition of the Delayed-Recognition task, which was subject to ICC analysis and detailed in Table 2 and Supplementary Table 1.

exhibits the least amount of activity with moderate-to-excellent reliability. Although the ICC values based on two or three fMRI sessions are not independent measures, these results converge to suggest that selectively vulnerable brain regions in aMCI exhibit reliable univariate BOLD activity during a visual selective-attention Delayed-Recognition task, although to a lesser extent than regions believed to be largely unaffected in aMCI.

Table 2
Whole brain univariate BOLD activity to face stimuli during the Delayed-Recognition task. Columns represent (in order): neural regions that exhibit significant BOLD activity, the number of significant voxels in that region, mean ICC of those voxels, mean range (upper–lower bound) of the 95% confidence interval (CI), the change in the CI range due to different number of fMRI sessions in the ICC estimate (three–two sessions), the number of voxels that exhibited moderate or better ICC values (ICC > 0.4), the percentage of voxels that exhibited moderate or better ICC values, and the rightmost column describes the percent change in the number of voxels with moderate or better ICC due to different number of fMRI sessions in the ICC estimate (three–two sessions).

	Region	# Voxels	Mean	Mean CI	ΔCI range	# Voxels	% Voxels	% Δ
		univariate	ICC	range	3–2 visits	ICC > 0.4	ICC > 0.4	3–2 visits
Frontal lobe	Precentral gyrus	1568	0.47	0.60	−0.25	1040	66%	12%
	Paracingulate gyrus	524	0.46	0.61	−0.27	324	62%	12%
	Juxtapositional lobule	459	0.42	0.63	−0.23	259	56%	−2%
	Middle frontal gyrus	1911	0.42	0.61	−0.27	1043	55%	15%
	Inferior frontal gyrus, pars opercularis	921	0.41	0.63	−0.27	463	50%	9%
	Superior frontal gyrus	666	0.38	0.63	−0.27	291	44%	6%
	Frontal pole	1180	0.38	0.63	−0.28	502	43%	12%
	Inferior frontal gyrus, pars triangularis	407	0.34	0.65	−0.24	146	36%	5%
	Frontal operculum cortex	209	0.31	0.65	−0.32	63	30%	14%
	Frontal orbital cortex	404	0.31	0.66	−0.25	109	27%	−7%
	Insular cortex	256	0.29	0.66	−0.30	65	25%	0%
	Cingulate gyrus, anterior	128	0.24	0.65	−0.28	18	14%	7%
	Total	8633	0.40	0.62	−0.27	4323	51%	10%
Temporal lobe	Temoral occipital fusiform	1146	0.54	0.56	−0.23	892	78%	9%
	Inferior temporal gyrus, temporooccipital	1204	0.43	0.62	−0.23	651	54%	4%
	Middle temporal gyrus, temporooccipital	597	0.38	0.63	−0.26	267	45%	−4%
	Temporal fusiform cortex, posterior	574	0.38	0.62	−0.23	230	40%	−2%
	Parahippocampal gyrus, posterior	235	0.35	0.65	−0.27	88	37%	5%
	Temporal pole	126	0.35	0.61	−0.26	40	32%	−1%
	Parahippocampal gyrus, anterior	519	0.25	0.66	−0.26	83	16%	−4%
	Total	4401	0.42	0.61	−0.24	2251	51%	2%
Parietal lobe	Precuneus cortex	524	0.63	0.50	−0.20	479	88%	7%
	Superior parietal lobule	917	0.50	0.59	−0.29	647	71%	23%
	Supramarginal gyrus, anterior	237	0.47	0.60	−0.25	155	65%	8%
	Supramarginal gyrus, posterior	712	0.45	0.61	−0.26	442	62%	13%
	Angular gyrus	270	0.44	0.61	−0.24	155	57%	8%
	Postcentral gyrus	320	0.27	0.65	−0.25	62	19%	9%
	Total	2980	0.48	0.59	−0.26	1940	65%	14%
Occipital lobe	Occipital pole	1688	0.64	0.49	−0.11	1529	91%	1%
	Lateral occipital cortex, inferior	2992	0.61	0.51	−0.20	2566	86%	6%
	Lateral occipital cortex, superior	3735	0.57	0.54	−0.19	3040	81%	3%
	Occipital fusiform gyrus	2462	0.58	0.52	−0.22	1957	79%	3%
	Intracalcarine cortex	531	0.47	0.60	−0.17	354	67%	−5%
	Lingual gyrus	1547	0.49	0.58	−0.21	974	63%	5%
	Total	12,955	0.58	0.53	−0.19	10,420	80%	3%
Sub-cortical	Caudate	122	0.37	0.64	−0.24	50	41%	3%
	Pallidum	123	0.32	0.66	−0.23	37	30%	−3%
	Amygdala	136	0.28	0.65	−0.32	29	21%	−2%
	Putamen	121	0.23	0.64	−0.29	15	12%	7%
	Total	502	0.30	0.65	−0.27	131	26%	1%

One-Back task

ICC on performance measures from the One-Back task was not assessed due to the simplicity of the task, small number of responses required, and difficulty discerning hits from false alarms (responses typically occurred after several stimuli had passed). However, ICC was calculated on the fMRI BOLD activity, as was done with the analysis of the Delayed Recognition task data.

Whole brain analysis

Univariate BOLD activity during the One-Back task mainly elicited activity in the occipital lobe when presented with face (Table 3) or scene (Supplementary Table 2) stimuli, and this is the region that displayed the highest amount of moderate-to-excellent reliability (faces: 52%, scenes: 61%). However, face stimuli evoked a large amount of activity throughout the frontal lobe (albeit to a lesser extent than the occipital lobe). Interestingly, only 28% of the frontal lobe activity to face stimuli exhibited moderate-to-excellent reliability scores, which was much lower than the extent of reliability in the occipital, temporal or parietal lobes. Similarly, the frontal and temporal lobes exhibited a large range of the 95% confidence intervals compared to occipital and parietal regions, further suggesting reliability of BOLD activity is less in the frontal and temporal regions compared to the occipital lobe.

When using the first two fMRI sessions to calculate ICC, the values change (relative to ICC from three fMRI sessions) based on stimulus type during the One-Back task. With two fMRI scans, ICC values decrease in every cluster of scene evoked BOLD activity, whereas ICC values from face evoked BOLD activity generally increased with fewer fMRI sessions. This ICC dependence on stimulus type is similar to that observed in the analysis of accuracy and ROI data, which most likely indicates that two fMRI sessions were not sufficient to resolve stable ICC measures. Moreover, the 95% confidence intervals were observed to increase when using only two, compared to three, fMRI sessions in the ICC estimate, suggesting additional fMRI sessions result in better reliability estimates. Nonetheless, ICC data from two and three fMRI sessions both indicate that the occipital lobe elicits the most reliable activity and the frontal lobe yields the least reliable activity during a visual One-Back task, but still exhibited fair reliability.

Comparison to Delayed Recognition task

To directly compare ICC measures from the two working memory tasks (i.e., Delayed Recognition and One-Back), the percent change of moderate-to-excellent reliability was calculated for each neural region in common to both tasks (Table 3 and Supplementary Table 2). Results show that regardless of stimulus type (faces or scenes), ICC values are generally higher in estimates from the Delayed Recognition task. Although the One-Back task did not evoke BOLD activity in the medial temporal lobe or in subcortical areas, ICC measures in task-related regions from both tasks (Delayed Recognition and One-Back) converge to show that the frontal lobe in aMCI individuals elicits activity with lower reliability than other neural regions during working memory tasks.

To better characterize the distribution of ICC values, Fig. 6 shows the percent of task-based univariate activity that exhibits a given minimum ICC value. Interestingly, scene stimuli for both working memory tasks appear to yield slightly higher reliability scores compared to face stimuli. It can be seen that when ICC is estimated from three fMRI sessions (Fig. 6A), over 60% of all Delayed Recognition univariate activity and over 35% of all One-Back activity yields at least a moderate (ICC > 0.4) estimate of reliability. Overall, the Delayed Recognition task elicits higher reliability measures than the One-Back task and using three fMRI sessions to estimate ICC (as opposed to two sessions, Fig. 6B) generally results in higher values (with the exception of the One-Back task with faces).

Reliability in non-task related regions

ICC values were calculated using all three fMRI sessions for voxels that did not exhibit significant task-evoked BOLD activity (i.e., voxels that were not included in the previous analyses). To assess reliability differences between task-related and non-task related BOLD activity, mean ICC values from non-significant regions were compared to each brain region listed in Tables 2 and 3, as well as Supplementary Tables 1 and 2. Results showed non-significant BOLD activity yields decreased ICC measures for each task, stimulus type, and every brain region assessed, except one. Specifically, non-significant activity during the Remember Faces condition of the Delayed Recognition task elicited an average ICC decrease of 0.14 (SEM = 0.02) compared to voxels of the same regions that exhibited significant BOLD activity (listed in Table 2). With the exception of the amygdala (ICC increased by 0.09), every region from Table 2 displayed less reliable BOLD activity in voxels that are not task-related. For non-significant BOLD activity during the Remember Scenes condition of the Delayed Recognition task, ICC values decreased on average by 0.12 (SEM = 0.02), where reliability in every

Table 3
Whole brain univariate BOLD activity when encoding face stimuli during the One-Back task. Column headings are similar to Table 2, with the addition of two columns that describe the change between the two tasks (Delayed Recognition (DR) – One Back task) in the 95% confidence interval range (middle column) as well as the change in the percent of voxels with moderate or better ICC values (rightmost column).

	Region	# Voxels univariate	Mean ICC	Mean CI range	ΔCI range 3–2 visits	Δ CI range DR-1Back	# Voxels ICC > 0.4	% Voxels ICC > 0.4	% Δ 3–2 visits	% Δ DR-1Back
Frontal lobe	Juxtapositional lobule	544	0.38	0.65	−0.22	−0.02	241	44%	−7%	12%
	Frontal operculum cortex	123	0.34	0.66	−0.30	−0.01	49	40%	14%	−10%
	Paracingulate gyrus	303	0.31	0.60	−0.14	0.01	107	35%	6%	27%
	Insular cortex	252	0.31	0.67	−0.28	−0.01	71	28%	−3%	−3%
	Cingulate gyrus, anterior	261	0.29	0.54	0.05	0.11	74	28%	−2%	−14%
	Precentral gyrus	615	0.27	0.66	−0.22	−0.06	134	22%	−21%	44%
	Inferior frontal gyrus, pars opercularis	346	0.26	0.67	−0.26	−0.02	63	18%	−15%	32%
	Superior frontal gyrus	166	0.27	0.67	−0.23	−0.04	29	17%	0%	27%
	Frontal pole	116	0.21	0.49	0.10	0.14	8	7%	3%	36%
	Total	2726	0.30	0.64	−0.19	0.01	776	28%	−7%	23%
Temporal lobe	Inferior temporal gyrus, temporooccipital	203	0.37	0.65	−0.22	−0.03	85	42%	−17%	12%
	Temoral occipital fusiform	744	0.34	0.65	−0.26	−0.09	283	38%	7%	40%
	Temporal fusiform cortex, posterior	130	0.25	0.68	−0.34	−0.06	19	15%	4%	25%
	Total	1077	0.33	0.65	−0.26	−0.06	387	36%	2%	15%
Parietal lobe	Superior parietal lobule	260	0.37	0.61	−0.07	−0.02	112	43%	1%	28%
Occipital lobe	Lateral occipital cortex, inferior	1745	0.45	0.61	−0.21	−0.10	1094	63%	4%	33%
	Occipital pole	279	0.44	0.62	−0.14	−0.13	174	62%	−19%	29%
	Occipital fusiform gyrus	840	0.38	0.65	−0.25	−0.13	392	47%	3%	32%
	Lateral occipital cortex, superior	829	0.31	0.58	−0.04	−0.04	246	30%	−19%	51%
	Total	3693	0.40	0.61	−0.18	−0.10	1906	52%	−3%	28%

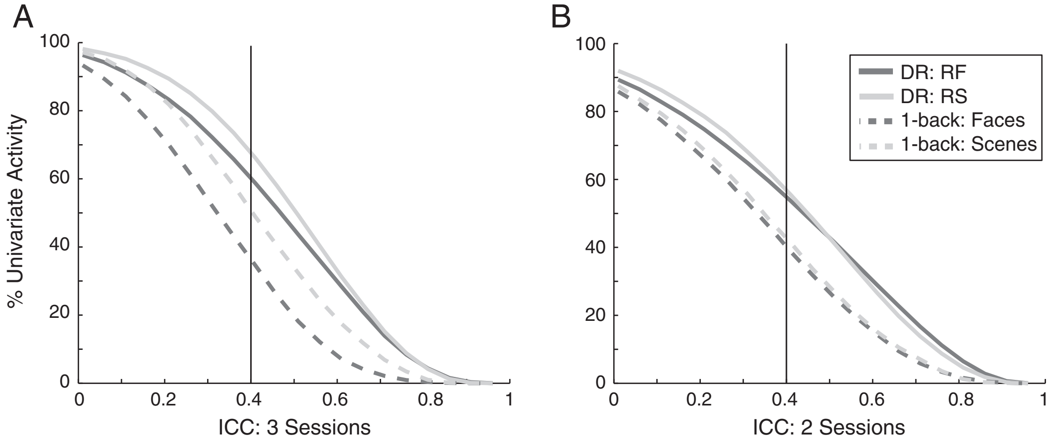

Fig. 6. Whole brain analysis depicting the percent of significant univariate BOLD activity with a specific magnitude of ICC or better.

region (listed in Supplementary Table 1) decreased with the exception of the intracalcarine cortex (ICC increased by 0.06). Similarly, during the One-Back task, reliability in voxels with non-significant BOLD activity decreased in every region (listed in Table 3 and Supplementary Table 2) with a mean ICC decrease of 0.10 (SEM = 0.02) and 0.12 (SEM = 0.06) during face and scene stimuli, respectively.

It could be argued that the ICC values for regions of non-task based activity are lower than those that exhibited task-based activity because of the much larger number of voxels included in this analysis. However, the occipital fusiform gyrus during the Delayed Recognition task elicited more voxels with significant (Remember Faces = 2462 voxels; Remember Scenes = 2385 voxels) than non-significant (Remember Faces = 1125 voxels; Remember Scenes = 1202 voxels) BOLD activity. Despite fewer voxels, mean ICC values from voxels with non-significant BOLD activity in the occipital fusiform gyrus resulted in lower reliability during both the Remember Faces (ICC decreased by 0.18) and Remember Scenes (ICC decreased by 0.13) conditions. This suggests that low ICC values in non-task related brain regions could not be fully attributed to differential sample sizes. Moreover, this underscores the utility of assessing reliability measures in task-evoked neural regions and implies that ROIs defined by anatomy may result in lower reliability measures than those driven by functional data.

Discussion

Little is known regarding the reliability of fMRI data in the aMCI population. Therefore, this study aimed to 1) assess reliability of BOLD activity in an aMCI population during two different visual working memory tasks and 2) explore the stability of ICC measurements using different numbers of fMRI sessions. Results indicated differential reliability measures based on the neural region assessed. Specifically, both tasks showed that the occipital lobe exhibited the highest ICC values such that the majority (52% to 82%) of significant task-based BOLD activity in the occipital lobe may be considered moderately reliable or better (i.e., ICC > 0.4). Subcortical structures exhibited the lowest reliability measures and BOLD activity in the frontal lobe demonstrated less reliability, relative to other cortical regions, with 28% to 60% of its BOLD activity exhibiting moderate or better ICC values. This suggests that the reliability of fMRI activity in posterior sensory regions may be more stable over visits than higher-level frontal association regions. Although these results indicate that neural regions known to decline in aMCI populations (i.e., frontal lobe, medial–temporal lobe, subcortical regions) yield the least reliable BOLD activity, it is important to highlight that the fMRI signal is still considered reliable in a population of aMCI participants who may actually suffer cognitive decline and brain atrophy between visits. Moreover, these results were observed when using two or three fMRI sessions in calculating ICC, although ICC estimates were generally higher and more stable when 3 fMRI sessions were used.

Face (FFA) and scene (PPA) ROI data yielded good reliability measurements (i.e., ICC > 0.6). Although the PPA exhibited similar ICC values between the Remember Faces and Remember Scenes conditions, the FFA displayed higher ICC values during the Remember Faces than Remember Scenes condition, supporting previous results indicating attended stimuli elicit greater reliability measures than ignored stimuli (Specht et al., 2003). Interestingly, both the FFA and PPA yielded better reliability scores than the hippocampus. It is possible that this reflects lower engagement of the hippocampus during the Delayed Recognition task, as activity in the hippocampus did not survive whole-brain multiple comparison corrections (hence its absence from Table 2 and Supplementary Table 1). Yet, reliability measures are not contingent on whether BOLD activity reaches low or high levels of significance (Caceres et al., 2009). It could be argued that the difference in the ROI findings may be methodological. In the present study, the FFA and PPA ROIs were identified using data from the One-Back task, which is specialized to elicit activity in these regions. The hippocampus ROI was identified at the group-level from data during the Delayed Recognition task. However, this methodological difference most likely favored the hippocampus ROI as it was defined by activity during the Delayed Recognition task, whereas activity in the FFA and PPA were not guaranteed to elicit significant BOLD activity during the task. We have provided evidence that non-significant BOLD activity elicits lower reliability measures, confirming previous results from healthy young adults (Aron et al., 2006), and as such, the FFA and PPA could be expected to elicit lower ICC measures than the hippocampus as activity in the FFA and PPA ROIs were not guaranteed to reach significance, whereas the hippocampal ROI was selected because of its significant (uncorrected) activity. Nonetheless, the hippocampus exhibited lower ICC values, which most likely cannot be attributed to ROI selection methodology. Thus, lower reliability of hippocampal activity more likely reflects degenerative changes of specific neural regions common to the aMCI population.

It could be argued that the differential reliability measures across the brain may be due to the assessment of data during the working memory encoding stage that could bias the occipital lobe for larger ICC values. However, this would not account for the larger reliability scores observed in the parietal lobe compared to frontal, temporal, or subcortical regions. Nonetheless, it is unclear if the variable measures of reliability across the brain are unique to the aMCI population. Some evidence indicates that healthy younger adults display consistent reliability across neural regions (Eaton NeuroImage 2008), which would suggest that deficits in aMCI are accompanied by diminished reliability of BOLD activity in susceptible brain areas. Yet, Clement and Belleville (2009) previously reported comparable reliability of fMRI BOLD activity between aMCI patients and healthy older adults. This would imply that decreased reliability that we observed in the frontal lobe, medial–temporal lobe, and subcortical areas are common in aged adults and not a product of abnormal neural decline. However, Clement and Belleville (2009) assessed reliability in two ways: 1) an overlap ratio that yields one value indicating whole-brain overlap of activity across fMRI sessions, and 2) they calculated ICC values from 8 ROIs (i.e., 5 PFC ROIs, 1 hippocampus, 1 precuneus, and 1 posterior cingulate cortex ROI), none of which were in occipital lobe and were not well represented in regions outside the frontal lobe. Thus, it is difficult to draw conclusions from their data regarding the distribution of reliability across the brain and whether it may differ between healthy aging and aMCI. As such, additional research will be required to fully delineate whether reduced reliability estimates in frontal lobe, medial–temporal lobe, and subcortical BOLD activity, compared to the rest of the brain, are due to normal aging or disease. Importantly, even though these regions may be susceptible to decline in functional integrity in aMCI, the reliability of these regions over 3 months was still fair-to-excellent suggesting that they are reasonably stable.

Interestingly, Clement and Belleville's (2009) memory-encoding task was most similar to the tasks reported here, but exhibited lower ICC values in their ROIs (ICC = 0.21 on average) than the most anatomically comparable regions assessed in the current study. This discrepancy most likely reflects the different methods used to identify the ROIs. Clement and Belleville (2009) utilized an anatomical atlas, whereas we selected ROIs based on task-evoked BOLD activity. As we have shown, non-task related regions yielded lower reliability estimates. Therefore, the previous report of lower ICC values is possibly a byproduct of assessing reliability in neural regions that are not involved in the task. This suggests that neuroimaging studies using anatomically guided ROIs might benefit from functionally driven ROIs based on an independent localizer task.

It should be noted that higher reliability measures were generally obtained during the Delayed Recognition task compared to the One-Back task. It could be argued that the lower reliability estimates during the One-Back task may be attributed to the simplicity of the task, thereby permitting mind-wandering or other non-task related neural processes that may lower the ICC values. However, performance (assessed online) was at ceiling, indicating that participants were vigilant during the task. Moreover, reliability estimates (via an overlap ratio) from tasks that have a motor component typically generate higher reliability measures (e.g., Miki et al., 2000; Rombouts et al., 1998) than non-motor tasks (e.g., Machielsen et al., 2000; Wagner et al., 2005). As the ICC values for the Delayed Recognition task were calculated during the encoding period (when no motor response was necessary), it could be hypothesized that the Delayed Recognition task would elicit lower reliability scores than the One-Back task. Yet, the Delayed Recognition task yielded greater reliability measures than the One-Back task, which more likely reflects the amount of power in the Delayed Recognition task data set due to the increased number of time-points. Indeed, fewer time-points lead to lower power estimates (Desmond and Glover, 2002), which would explain the lower reliability estimates during the One-Back task. In support of this, the 95% confidence intervals of the ICC estimates were generally larger during the One-Back task. Therefore, these results underscore the necessity to design an adequately powered experiment with a sufficient number of time-points in order to increase the reliability of the data.

Along the same lines, ICC values were generally higher and the 95% confidence intervals were smaller when ICC was calculated using three fMRI sessions as opposed to using two sessions. The comparisons between ICC values based on two or three fMRI sessions are not independent measures, and the general increase in ICC may partially reflect a biasing due to increased power in the ICC calculation. However, ICC values were only marginally higher and did not always increase when using three fMRI sessions. As such, ICC results using three fMRI sessions appeared to yield more stable (or reliable) reliability estimates. For example, reliability of accuracy was expected to be comparable between the Remember Faces and Remember Scenes conditions during the Delayed Recognition task, which was only observed when calculating ICC based on three (and not two) fMRI sessions. Additionally, the FFA and PPA ROIs were expected to display higher reliability measures when face and scene stimuli were attended, respectively, relative to ignored stimuli (Specht et al., 2003). This was observed in the FFA (i.e., Remember Faces > Remember Scenes) when ICC was calculated from two or three fMRI sessions. Yet, the exact opposite was observed in the PPA (i.e., expected: Remember Scenes > Remember Faces, observed Remember Faces > Remember Scenes) when ICC was based on two fMRI sessions. Interestingly, comparable reliability estimates were obtained in the PPA (i.e., Remember Scenes = Remember Faces) when ICC was calculated from three sessions. Although ICC from two sessions did not exhibit the hypothesized results, the addition of one extra fMRI session in the ICC calculation resulted in a trend in the expected direction. Moreover, the 95% confidence intervals were consistently smaller when calculating ICC based on three, compared to two, fMRI sessions. Overall, these data seem to indicate that additional fMRI sessions elicit more stable measures of reliability.

For a more pragmatic interpretation of the results, we must decide how good ICC values should be for them to be considered good enough. The interpretation of ICC values is straightforward in that 1.0 depicts near perfect agreement between test and retest sessions whereas 0.0 indicates no agreement. However, there is no current consensus on acceptable ICC values for fMRI data. In accordance with previous reliability studies of fMRI data, we categorized ICC values greater than 0.4 as moderate or better (Ances et al., 2011; Eaton et al., 2008; Guo et al., 2012), which we interpret as a reasonable level of reliability. Thus, Tables 2 and 3 as well as Supplementary Tables 1 and 2 include data on the distribution of voxels with moderate reliability or better, which may serve future studies using a Delayed-Recognition or One-Back task. Specifically, these working memory paradigms elicited the most stable measures of BOLD activity in the occipital cortex. As such, future studies assessing visual functions in aMCI through a region of interest approach may be better served by examining occipital cortex. Moreover, future assessments of reliability will benefit from additional fMRI sessions when calculating ICC, although it remains unclear how many (or whether) extra sessions will be required before ICC estimates become asymptotic. In practice, it may not be possible to acquire more than two sessions of data for a reliability analysis, in which case, increasing the number of experimental trials per session may result in better reliability estimates.

Assessing reliability of performance data was not a main goal of this study. However, it is important to note that response time data during the Delayed Recognition task was a more reliable performance measure than memory accuracy. Thus, future assessments of the aMCI population using delayed recognition paradigms will benefit from analyses that incorporate response time data, and not focus purely on accuracy as the sole performance metric. Given that accuracy and response time are thought to assess different aspects of memory (Macleod and Nelson, 1984), this result underscores the importance of utilizing response time data as a metric for memory processes, especially in aMCI populations where memory decline is often the behavioral outcome of interest.

Conclusions

The results of this study indicate that additional data results in better estimates of reliability as well as more reliable BOLD activity. Thus, future studies on fMRI reliability will benefit from increasing the number of sessions used to calculate reliability metrics. Additionally, neuroimaging studies that are powered by a sufficient number of trials will yield results that are easier to replicate. These findings could aid in the design of future clinical trials interested in examining fMRI outcomes. Based on the comparison of 2 working memory tasks, the findings suggest that a delayed recognition task with a minimum of 30 trials per condition would produce a moderate or better reliability in many regions susceptible to changes in aMCI. This consideration is particularly relevant given the demonstrated risk of decreased reliability in regions with selectively vulnerability. Overall, longitudinal fMRI studies of cognitive function, such as memory processing, can be successfully utilized in older adults with aMCI and the noted recommendations should be considered to ensure adequate statistical power.

Acknowledgments

This work was supported by Pfizer/Eisai [A.G.], NIH grants R01MH096861 [A.G., T.Z], and K01AG034175 [J.P.]. We would also like to thank Ian Greenhouse and Anne Berry for their assistance in data collection.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2013.08.063.

References

Ances, B., Vaida, F., Ellis, R., Buxton, R., 2011. Test–retest stability of calibrated BOLD-fMRI in HIV- and HIV plus subjects. Neuroimage 54, 2156–2162.

Aron, A.R., Gluck, M.A., Poldrack, R.A., 2006. Long-term test–retest reliability of functional MRI in a classification learning task. Neuroimage 29, 1000–1006.

Atri, A., O'Brien, J.L., Sreenivasan, A., Rastegar, S., Salisbury, S., DeLuca, A.N., O'Keefe, K.M., LaViolette, P.S., Rentz, D.M., Locascio, J.J., Sperling, R.A., 2011. Test–retest reliability of memory task functional magnetic resonance imaging in Alzheimer disease clinical trials. Arch. Neurol. 68, 599–606.

Caceres, A., Hall, D.L., Zelaya, F.O., Williams, S.C.R., Mehta, M.A., 2009. Measuring fMRI reliability with the intra-class correlation coefficient. Neuroimage 45, 758–768.

Census-Bureau, U.S., 2008. Projections for the population by age and sex for the United States: 2010 to 2050. Population-Division, Washington, D.C.

Chen, E.E., Small, S.L., 2007. Test–retest reliability in fMRI of language: group and task effects. Brain Lang. 102, 176–185.

Cherubini, A., Peran, P., Spoletini, I., Di Paola, M., Di Iulio, F., Hagberg, G.E., Sancesario, G., Gianni, W., Bossu, P., Caltagirone, C., Sabatini, U., Spalletta, G., 2010. Combined volumetry and DTI in subcortical structures of mild cognitive impairment and Alzheimer's disease patients. J. Alzheimers Dis. 19, 1273–1282.

Clement, F., Belleville, S., 2009. Test–retest reliability of fMRI verbal episodic memory paradigms in healthy older adults and in persons with mild cognitive impairment. Hum. Brain Mapp. 30, 4033–4047.

De Vogelaere, F., Santens, P., Achten, E., Boon, P., Vingerhoets, G., 2012. Altered default-mode network activation in mild cognitive impairment compared with healthy aging. Neuroradiology 54, 1195–1206.

Delis, D.C., Kramer, J.H., Kaplan, E., Ober, B.A., 2000. California Verbal Learning Test, second ed. The Psychological Corporation, San Antonio, TX.

Delis, D., Kaplan, E.B., Kramer, J., 2001. The Delis–Kaplan Executive Function System. The Psychological Corporation, San Antonio, TX.

Desikan, R.S., Segonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31, 968–980.

Desmond, J.E., Glover, G.H., 2002. Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. J. Neurosci. Methods 118, 115–128.

Dickerson, B.C., Sperling, R.A., 2008. Functional abnormalities of the medial temporal lobe memory system in mild cognitive impairment and Alzheimer's disease: insights from functional MRI studies. Neuropsychologia 46, 1624–1635.

Eaton, K.P., Szaflarski, J.P., Altaye, M., Ball, A.L., Kissela, B.M., Banks, C., Holland, S.K., 2008. Reliability of fMRI for studies of language in post-stroke aphasia subjects. Neuroimage 41, 311–322.

Epstein, R., Kanwisher, N., 1998. A cortical representation of the local visual environment. Nature 392, 598–601.

Fernandez, G., Specht, K., Weis, S., Tendolkar, I., Reuber, M., Fell, J., Klaver, P., Ruhlmann, J., Reul, J., Elger, C.E., 2003. Intrasubject reproducibility of presurgical language lateralization and mapping using fMRI. Neurology 60, 969–975.

Ferreira, L.K., Diniz, B.S., Forlenza, O.V., Busatto, G.F., Zanetti, M.V., 2011. Neurostructural predictors of Alzheimer's disease: a meta-analysis of VBM studies. Neurobiol. Aging 32, 1733–1741.

Gazzaley, A., Cooney, J.W., Rissman, J., D'Esposito, M., 2005. Top-down suppression deficit underlies working memory impairment in normal aging. Nat. Neurosci. 8, 1298–1300.

Gazzaley, A., Clapp, W., Kelley, J., McEvoy, K., Knight, R., D'Esposito, M., 2008. Age-related top-down suppression deficit in the early stages of cortical visual memory processing. Proc. Natl. Acad. Sci. U. S. A. 105, 13122–13126.

Gomez-Tortosa, E., Mahillo-Fernandez, I., Guerrero, R., Montoya, J., Alonso, A., Sainz, M.J., 2012. Outcome of mild cognitive impairment comparing early memory profiles. Am. J. Geriatr. Psychiatry 20, 827–835.

Guo, C.C., Kurth, F., Zhou, J., Mayer, E.A., Eickhoff, S.B., Kramer, J.H., Seeley, W.W., 2012. One-year test–retest reliability of intrinsic connectivity network fMRI in older adults. Neuroimage 61, 1471–1483.

Kanwisher, N., McDermott, J., Chun, M.M., 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. 17, 4302–4311.

Kimberley, T.J., Khandekar, G., Borich, M., 2008. fMRI reliability in subjects with stroke. Exp. Brain Res. 186, 183–190.

Kramer, J.H., Jurik, J., Sha, S.J., Rankin, K.P., Rosen, H.J., Johnson, J.K., Miller, B.L., 2003. Distinctive neuropsychological patterns in frontotemporal dementia, semantic dementia, and Alzheimer disease. Cogn. Behav. Neurol. 16, 211–218.

Kurland, J., Naeser, M.A., Baker, E.H., Doron, K., Martin, P.I., Seekins, H.E., Bogdan, A., Renshaw, P., Yurgelun-Todd, D., 2004. Test–retest reliability of fMRI during nonverbal semantic decisions in moderate-severe nonfluent aphasia patients. Behav. Neurol. 15, 87–97.

Machielsen, W.C.M., Rombouts, S., Barkhof, F., Scheltens, P., Witter, M.P., 2000. fMRI of visual encoding: reproducibility of activation. Hum. Brain Mapp. 9, 156–164.

Macleod, C.M., Nelson, T.O., 1984. Response latency and response accuracy as measures of memory. Acta Psychol. 57, 215–235.

Manoach, D.S., Halpern, E.F., Kramer, T.S., Chang, Y.C., Goff, D.C., Rauch, S.L., Kennedy, D.N., Gollub, R.L., 2001. Test–retest reliability of a functional MRI working memory paradigm in normal and schizophrenic subjects. Am. J. Psychiatry 158, 955–958.

Miki, A., Raz, J., van Erp, T.G.M., Liu, C.S.J., Haselgrove, J.C., Liu, G.T., 2000. Reproducibility of visual activation in functional MR imaging and effects of postprocessing. Am. J. Neuroradiol. 21, 910–915.

Missonnier, P., Gold, G., Fazio-Costa, L., Michel, J.P., Mulligan, R., Michon, A.S., Ibanez, V., Giannakopoulos, P., 2005. Early event-related potential changes during working memory activation predict rapid decline in mild cognitive impairment. J. Gerontol. A Biol. Sci. Med. Sci. 60, 660–666.

Missonnier, P., Gold, G., Herrmann, F.R., Fazio-Costa, L., Michel, J.P., Deiber, M.P., Michon, A., Giannakopoulos, P., 2006. Decreased theta event-related synchronization during working memory activation is associated with progressive mild cognitive impairment. Dement. Geriatr. Cogn. Disord. 22, 250–259.

Pa, J., Berry, A.S., Compagnone, M., Boccanfuso, J., Greenhouse, I., Rubens, M.T., Johnson, J.K., Gazzaley, A., 2013. Cholinergic modulation of functional network connectivity associated with attention and memory in older adults with mild memory deficits. Ann. Neurol. 73 (6), 762–773.

Petersen, R.C., 1995. Normal aging, mild cognitive impairment, and early Alzheimer's disease. Neurologist 1, 326–344.

Petersen, R.C., 2000. Aging, mild cognitive impairment, and Alzheimer's disease. Neurol. Clin. 18, 789.

Possin, K.L., Laluz, V.R., Alcantar, O.Z., Miller, B.L., Kramer, J.H., 2011. Distinct neuroanatomical substrates and cognitive mechanisms of figure copy performance in Alzheimer's disease and behavioral variant frontotemporal dementia. Neuropsychologia 49, 43–48.

Rombouts, S., Barkhof, F., Hoogenraad, F.G.C., Sprenger, M., Scheltens, P., 1998. Within-subject reproducibility of visual activation patterns with functional magnetic resonance imaging using multislice echo planar imaging. Magn. Reson. Imaging 16, 105–113.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations — Uses in assessing rater reliability. Psychol. Bull. 86, 420–428.

Specht, K., Willmes, K., Shah, N.J., Jancke, L., 2003. Assessment of reliability in functional imaging studies. J. Magn. Reson. Imaging 17, 463–471.

Squire, L.R., Stark, C.E.L., Clark, R.E., 2004. The medial temporal lobe. Annu. Rev. Neurosci. 27, 279–306.

Stroop, J.R., 1935. Studies of interference in serial verbal reactions. J. Exp. Psychol. 18, 643.

Tombaugh, T.N., 2004. Trail making Test A and B: normative data stratified by age and education. Arch. Clin. Neuropsychol. 19, 203–214.

Wagner, K., Frings, L., Quiske, A., Unterrainer, J., Schwarzwald, R., Spreer, J., Halsband, U., Schulze-Bonhage, A., 2005. The reliability of fMRI activations in the medial temporal lobes in a verbal episodic memory task. Neuroimage 28, 122–131.

Wechsler, D., 1981. Wechsler Adult Intelligence Scale—Revised manual. The Psychological Corporation, New York, NY.

Wechsler, D., 1987. Wechsler Memory Scale—Revised manual. The Psychological Corporation, San Antonio, TX.

Yang, J., Pan, P.L., Song, W., Huang, R., Li, J.P., Chen, K., Gong, Q.Y., Zhong, J.G., Shi, H.C., Shang, H.F., 2012. Voxelwise meta-analysis of gray matter anomalies in Alzheimer's disease and mild cognitive impairment using anatomic likelihood estimation. J. Neurol. Sci. 316, 21–29.