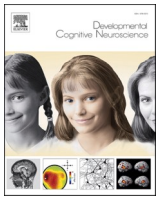
Developmental Cognitive Neuroscience 52 (2021) 101037

Contents lists available at ScienceDirect

Developmental Cognitive Neuroscience

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



Reliability of subsequent memory effects in children and adults: The good, the bad, and the hopeful

Lingfei Tang [a,b], Qijing Yu [a,b], Roya Homayouni [a,b], Kelsey L. Canada [a], Qin Yin [a,b], Jessica S. Damoiseaux [a,b], Noa Ofen [a,b,*]

[a] Institute of Gerontology, Wayne State University, Detroit, MI, United States
[b] Department of Psychology, Wayne State University, Detroit, MI, United States

ARTICLE INFO

Keywords:
FMRI
Reliability
Reproducibility
Subsequent memory
Development
Test-retest

ABSTRACT

Functional MRI (fMRI) is a key tool for investigating neural underpinnings of cognitive development. Yet, in recent years, the reliability of fMRI effects has come into question and with it, the feasibility of using task-based fMRI to identify developmental changes related to cognition. Here, we investigated the reliability of task-based fMRI activations with a widely used subsequent memory paradigm using two developmental samples: a cross-sectional sample (n = 85, age 8–25 years) and a test-retest sample (n = 24, one-month follow up, age 8–20 years). In the large cross-sectional sample, we found good to excellent group-level reliability when assessing activation patterns related to the encoding task and subsequent memory effects. In the test-retest sample, while group-level reliability was excellent, the consistency of activation patterns within individuals was low, particularly for subsequent memory effects. We observed consistent activation patterns in frontal, parietal, and occipital cortices, but comparatively lower test-retest reliability in subcortical regions and the hippocampus. Together, these findings highlight the limitations of interpreting task-based fMRI effects and the importance of incorporating reliability analyses in developmental studies. Leveraging larger and densely collected longitudinal data may help contribute to increased reproducibility and the accumulation of knowledge in developmental sciences.

1. Introduction

The ability to form vivid memories of past events is crucial in guiding everyday decisions and social interactions. Dramatic improvements in memory functioning from childhood to adulthood continue to draw efforts to elucidate the neural mechanisms that support memory development. With the advent of functional MRI (fMRI) as a neuroimaging technique, scientists have been able to characterize the neural substrates of memory by measuring changes in the blood oxygen-level dependent (BOLD) signal in the brain while participants complete a memory task in the scanner. Accumulating fMRI evidence continues to enrich our understanding of how changes in functional activations of specific regions contribute to memory development. However, there is recent and growing appreciation of the many methodological factors that can influence the ability to adequately interpret fMRI activations in order to characterize true developmental effects.

Examining developmental effects reported in the extant fMRI studies on memory development evinces inconsistent patterns of age differences, likely reflecting the limited ability to characterize true developmental trajectories with this method. To illustrate, we compare findings for two brain regions that are important for memory, the prefrontal cortex (PFC) and the hippocampus (Scoville and Milner, 1957; Shimamura, 1995). While previous research has identified convergent developmental effects in the PFC, with continued increases in the levels of activation and deactivation from childhood to adulthood (Ghetti and Bunge, 2012; Güler and Thomas, 2013; Ofen, 2012; Tang et al., 2018), consistent patterns have not been identified for developmental effects in the hippocampus. The hippocampus and its adjacent cortices show age invariance in their activation based on a number of cross-sectional studies (Güler and Thomas, 2013; Ofen et al., 2012, 2007; Shing et al., 2016; Tang et al., 2020), but other studies have noted age-related increases (e.g., DeMaster et al., 2013; Ghetti et al., 2010; Paz-Alonso et al., 2008) or decreases (Maril et al., 2010) in the contribution of these regions to memory. The inconsistent findings may reflect differences

* Correspondence to: Wayne State University, Institute of Gerontology, 87 E. Ferry St., Detroit, MI 48202, United States.
E-mail address: noa.ofen@wayne.edu (N. Ofen).

https://doi.org/10.1016/j.dcn.2021.101037
Received 18 April 2021; Received in revised form 27 October 2021; Accepted 16 November 2021
Available online 17 November 2021

between studies in methodology, research paradigms, and/or sample characteristics. For example, some of the studies mentioned above adopted intentional encoding (e.g., DeMaster et al., 2013; Ofen et al., 2007; Tang et al., 2020), while others used an incidental encoding task (Ghetti et al., 2010). A recent longitudinal study by Nolden et al. (2021) showed that children as young as 5 years old utilized the hippocampus during intentional memory encoding, even though this effect did not differ between the group that remained in kindergarten and the group that went on to elementary school. These points aside, the discrepancies in hippocampal age effects nonetheless raise questions regarding the confidence in using fMRI to understand memory development.

More broadly, the reliability of task-based fMRI in general has come into question during recent years. A growing body of literature show that the measurement reliability of activation patterns in fMRI tasks that are commonly used and widely validated remains modest at best, and the reliability is particularly low in subcortical regions in the brain (Elliott et al., 2020; Bennett and Miller, 2013). Suboptimal reliability, reflecting high measurement error, could result in difficulty extracting signal from noise and, in the context of developmental research, hinder our ability to detect true development effects. While a few studies have commented on the reliability of memory paradigms in healthy and clinical adults (Bennett and Miller, 2013; Brandt et al., 2013; Clément and Belleville, 2009; Harrington et al., 2006; Putcha et al., 2011; Towgood et al., 2015), the reliability of assessing neural correlates of memory in typically developing populations remains unknown. Systematic evaluations of the strength and limitations of these memory paradigms, ideally prior to conducting targeted developmental analyses, become paramount.

When investigating the neural correlates underpinning cognitive development, task selection is usually motivated by a researcher's notion of what task is best suited for measuring behavior that is reflective of the desired cognitive construct. However, the reliability of the chosen paradigm is often neglected and can heavily influence the estimation of reliability in task-based fMRI, even when they are targeting the same construct (Bossier et al., 2020; Gorgolewski et al., 2013; Plichta et al., 2012; Turner et al., 2018; Vetter et al., 2017). Therefore, we argue that the reliability of a paradigm should be established first (e.g., Boenniger et al., 2021). One approach to examining paradigm reliability is to assess the internal consistency of the experimental task with a split-half method, where all trials of the paradigm divided into two halves and behavioral and fMRI results between the two halves are compared. Good consistency between the two halves would indicate that the paradigm itself is reliable.

In terms of selecting a paradigm targeting neural correlates of memory, perhaps the most long-standing and widely used paradigm in fMRI studies is the subsequent memory paradigm. In this paradigm, participants actively study stimuli while being scanned, and memory-related activation and deactivation can be computed based on comparing studied stimuli subsequently remembered to those subsequently forgotten. This paradigm has been used to identify subsequent memory effects (SME), that is, differential BOLD signals for items later remembered versus later forgotten. Previous research has implicated key regions involved in memory encoding, such as the MTL and PFC (Paller and Wagner, 2002, for a review, see Kim, 2011). Moreover, this paradigm has been applied to participants from wide age ranges, including children and older adults (de Chastelaine et al., 2011; Ghetti and Bunge, 2012; Ofen, 2012; Shing et al., 2010). The subsequent memory paradigm therefore serves as an ideal candidate for reliability analyses and is utilized in the current study. While few previous studies examined the reliability of memory tasks, they either focused on memory retrieval (Bennett and Miller, 2013) or examined memory encoding with a whole encoding block (e.g. Brandt et al., 2013; Clément and Belleville, 2009). One previous study (Putcha et al., 2011) implemented a subsequent memory paradigm, but did not examine SME (Hit vs. Miss) due to design complications ("limited jitter and small number of miss trials"), and instead focused on other contrasts not directly related to subsequent memory (e.g., High Confidence Hit vs. Repeated Pair). In this study, we focus specifically on the reliability of encoding and memory effects with a subsequent memory paradigm.

In addition to the experimental paradigm, the choice of fMRI contrast can also drastically influence the estimation of reliability. Previous research has shown that contrasts with conditions against an implicit baseline (e.g., viewing faces or viewing scenes) generally yield more reliable activations compared to contrasts between two independent conditions (e.g., viewing scenes > viewing faces, see Aron et al., 2006; Bennett and Miller, 2013; Vetter et al., 2017). In conducting reliability experiments, researchers are often at liberty to report specific fMRI contrasts. However, due to the apparent differences in reliability by fMRI contrasts, the inconsistent choice of contrast makes the comparisons of reliability measures between different studies difficult. Therefore, when examining the reliability of a subsequent memory paradigm, we report both the contrast relating to the encoding task (against the implicit baseline) and the contrast targeting specifically the memory construct that compares two conditions (remembered > forgotten).

To establish the reliability of behavioral and fMRI contrasts, one approach is to assess the consistency of effects between subsamples drawn from the same dataset (Fröhner et al., 2019). Group-level results of individual contrasts can be calculated for each subsample, and between-subject reliability can be computed by comparing results from these subsamples. A strength of this approach is that researchers can utilize a cross-sectional sample to assess measurement reliability, with larger samples providing greater power to establish an estimate of reliability. However, there are clear limitations in drawing inferences about reliability solely from group-level analyses, as this approach cannot isolate variances due to intra-individual differences from variances due to inter-individual differences. Previous research has shown marked differences in task-based fMRI reliability when comparing group-level results and individual-level results (Raemaekers et al., 2007; Van Den Bulk et al., 2013; Vetter et al., 2017). Further, individual-level analyses can shed light on the consistency of measures for the same person over time, a prerequisite for the longitudinal design in developmental studies. Consequently, individual-level reliability must also be considered, and researchers can collect data from multiple scans of the same individual over a relatively short delay to examine test-retest reliability. The activation maps generated for each person at multiple time-points are then compared to derive reliability measures.

In this study, we investigated the reliability of fMRI measures based on a commonly used subsequent memory paradigm in children and young adults. We utilized two developmental samples: (1) a cross-sectional sample of 85 participants scanned once and (2) a test-retest sample of 24 participants scanned at two time-points, one month apart. We measured reliability in several complementary ways. First, we determined the internal consistency of the paradigm by comparing group-level activation maps computed based on two halves of encoding trials. Second, we assessed the group-level reliability by comparing group results of two non-overlapping subsamples (40 participants for each subsample). Third, we examined both group- and individual-level reliability with the longitudinal test-retest sample. Each of the three main objectives was assessed for activation maps computed based on contrasting (1) encoding trials, regardless of subsequent memory, compared to implicit baseline (Encoding All Trials), and (2) encoding of stimuli subsequently recognized compared to those subsequently forgotten (SME, Subsequent Hit vs. Subsequent Miss trials). We predicted that the reliability of fMRI measures will differ by several key factors: assessing on the group- versus individual-level, choice of contrast, and regions in the brain. Last, we explored whether reliability differs by age, as suggested in prior work (Koolschijn et al., 2011). Understanding how these factors influence the degree of measurement reliability in task-based fMRI studies and developmental samples is critical for designing studies and drawing adequate conclusions on the neural correlates of memory development.

2. Materials and methods

2.1. Participants

Two samples of participants were recruited from the Metro Detroit area. All participants were right-handed, had normal or corrected-to-normal vision, were not claustrophobic, and had no known history of psychiatric or neurological disorders. They provided informed consent or assent as per a Wayne State University IRB-approved protocol and monetarily compensated for their time in this study. For the cross-sectional sample, we included 85 participants (ages 8–25 years, mean ± SD: 16.51 ± 4.73 years), as part of the ongoing data collection reported in previous publications (Tang et al., 2020, 2018). Unrelated to the current study, some of the participants in the cross-sectional sample has longitudinal (> 2 years) follow-ups, in which case data from only the first visit was included for each participant. For the test-retest sample, we collected data from 24 new participants (ages 8–20 years, 13.31 ± 3.11 years), who visited the imaging center twice (age in years at Visit 1: 13.27 ± 3.14; Visit 2: 13.35 ± 3.11), which took place approximately one month apart (30.26 ± 3.04 days between visits). Apart from the 85 participants from the cross-sectional sample and the 24 from the test-retest sample, an additional 22 participants (18 participants from the cross-sectional sample and 4 participants from the test-retest sample) were excluded from the study due to incomplete data, technical difficulties, or excessive motion (mean framewise displacement > 1 mm). Importantly, to ensure sufficient power for fMRI analysis, we included participants with at least 15 Hit or Miss trials for each condition across all samples and visits. There was no overlap in participants between the two datasets.

2.2. Subsequent memory paradigm stimuli and behavioral measures

The subsequent memory paradigm, described in detail in previous publications (Ofen et al., 2007; Tang et al., 2020, 2018) was administered to all participants. In brief, participants studied 120 pictures of indoor and outdoor scenes in the scanner. They were instructed to respond with a button press whether each picture depicted an indoor or outdoor scene (see Fig. 1A for an illustration of the paradigm). In addition to making an indoor/outdoor decision, participants were also explicitly instructed to memorize the scenes for a subsequent recognition test. Each scene was presented for 3 s, followed by a 0.5 s fixation cross and a variable inter-trial interval ranging from 0 to 12 s (negative exponential distribution). Variable inter-trial interval was used to optimize fMRI measurement (sequence determined using optseq2, http://surfer.nmr.mgh.harvard.edu/optseq/). Scenes were presented in 3 consecutive runs, each including 40 scenes. Each run lasted for 3 min and 54 s. Approximately 15 min after the completion of the MRI session, participants completed a self-paced recognition test outside the scanner with all 120 studied scenes presented intermixed with 80 foil scenes.

Scenes were drawn from a stimuli set of 600 images that were organized into 15 lists of 40 each. Half of the scenes in a list depicted indoor scenes and the other half depicted outdoor scenes. In addition, each list was balanced by scene complexity, which was determined by the number of unique items in the scene (as described in details in Chai et al., 2010), although it is not the focus of the current study. Each participant studied a predetermined set of 3 lists of the 15 potential lists during encoding. During recognition, participants were tested on the 120 images from the 3 studied lists intermixed with 80 images from a new set of 2 predetermined lists. Pseudorandomized and counterbalanced list assignment was used across participants. Participants in the test-retest reliability study received different lists of stimuli for each visit. Encoding task response accuracy (indoor/outdoor) and reaction

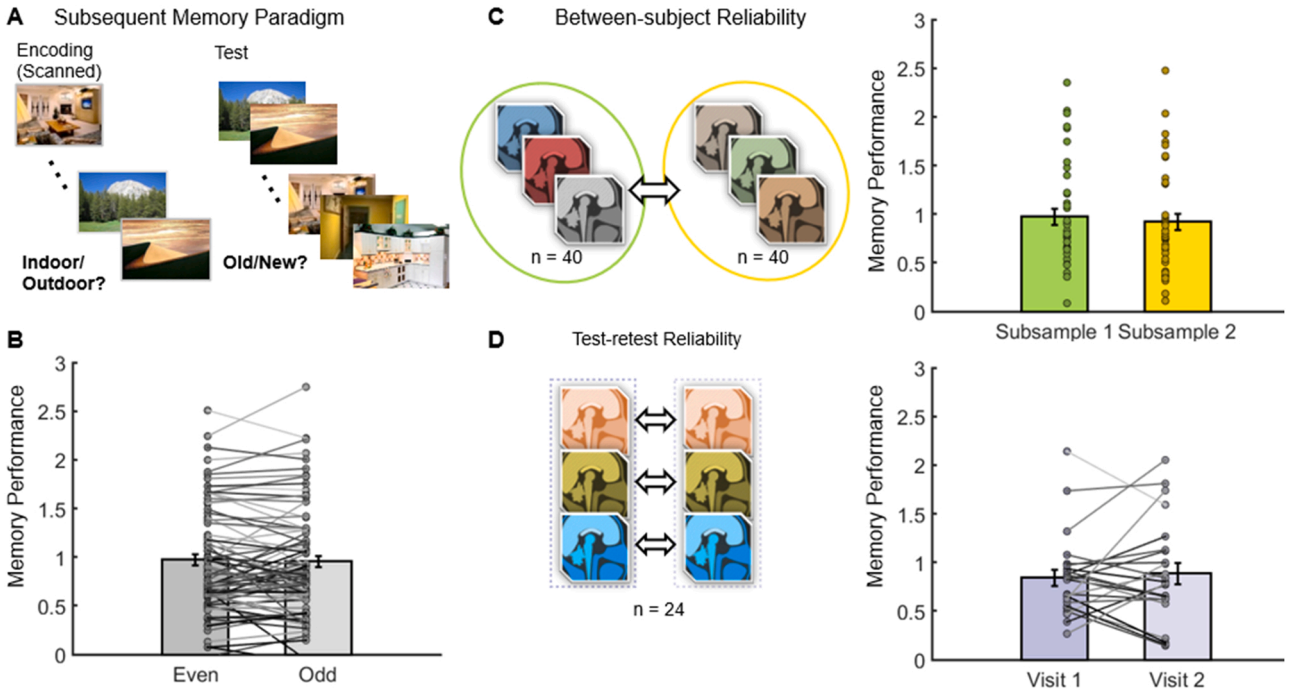


Fig. 1. Subsequent memory paradigm and the reliability of behavioral measures. A. Participants encoded a total of 120 indoor and outdoor scenes while being scanned and were administered a recognition test with the 120 studied scenes, intermixed with 80 foils, shortly after the scan session. Memory performance was calculated as the sensitivity index (d') considering Hit rate corrected by False Alarm rate. B. Memory performance calculated separately for even and odd trials indicated excellent internal reliability of the paradigm, with comparable group-level d' based on even and odd trials (Even trials: 0.97 ± 0.54, Odd trials: 0.95 ± 0.56, ICC = .89). C. Memory performance calculated separately for two representative subsamples of the large cross-sectional dataset indicate good between-subject reliability evidenced by comparable d' (Subsample 1: 0.97 ± 0.55, Subsample 2: 0.92 ± 0.54, $p = .67$). D. Memory performance calculated for each of the visits in the test-retest sample evinced good reliability in d' between visits (Visit 1: 0.84 ± 0.42, Visit 2: 0.88 ± 0.52; ICC = .62). In B and D, greyscale lines denote individual participants, and the shades of the lines represent relative participant age within the group (darker lines represent younger participants).

times (RTs) for the response were recorded for each encoding trial. Based on responses given during the recognition test, encoding trials were back labeled as subsequent Hits (subsequently recognized as old) or subsequent Misses (subsequently judged as new). Recognition accuracy was calculated as d' reflecting Hit rate adjusted by the rate participants falsely endorsing foils (False Alarms, FA).

2.3. MRI data acquisition

MRI data were acquired in a 3T Siemens Verio scanner at the Harper University Hospital in Detroit, Michigan. T1-weighted whole-brain structural images were acquired using an MPRAGE sequence [192 sagittal slices, repetition time (TR) = 2200 ms, echo time (TE) = 4.26 ms, flip angle = 9°, field of view = 256 mm, 192 × 256 voxels, and voxel size = 1 mm × 0.5 mm × 1 mm]. After the structural scan, functional images were acquired using a T2*-weighted gradient-echo sequence. Thirty sagittal slices were collected parallel to the AC-PC plane (TR = 2000 ms, TE = 30 ms, flip angle = 90º, effective voxel size = 3.125 mm × 3.125 mm × 4.8 mm). At each visit, participants were scanned for three consecutive functional runs while engaging in a subsequent memory paradigm. Each functional run consisted of 118 volume acquisitions. All participants included in this study underwent one structural scan and three consecutive functional runs for each visit, with the same MRI sequences.

2.4. Reliability quantification of behavioral measures

To determine the reliability in behavioral measures, the intra-class correlation, or ICC was used (Bennett and Miller, 2010; Caceres et al., 2009; Herting et al., 2018). We quantified the reliability using ICC(2), a two-way random model with absolute agreement, which measures the ratio of between-subject variance out of total variance, generating a value between 0 and 1 (theoretically). The ICC can be used to examine the consistency in observed values between samples and visits and allows for comparison between different studies. Conventionally, ICC less than .40 indicates poor reliability; ICC between .41 to .59 indicates fair reliability; ICC between .60 and .74 indicates good reliability; ICC between .75 and 1 indicates excellent reliability (Cicchetti, 2001). ICCs are reported for behavioral measures for the three main objectives, namely: (1) assessing the consistency in effects based on split half of trials, indicating the internal reliability of the study paradigm, (2) assessing between-subject group-level reliability in the cross-sectional sample, and (3) assessing group- and individual-level reliability in the test-retest sample.

2.5. Reliability analysis of behavioral measures

Encoding trials were categorized as subsequent Hit or subsequent Miss based on recognition memory outcome, whereas foils used during recognition were categorized as False Alarm (FA, incorrectly identified as old) or Correct Rejection (correctly identified as new). Memory performance was measured by the sensitivity index d' [z(Hit rate) – z(FA rate)] and average reaction time (RT) was calculated separately for subsequent Hit and Miss trials. Mean and SD statistics were reported for behavioral data, including Hit rate, FA rate, d', and RT.

Consistency and reliability of the behavioral effects were assessed for the three main objectives. First, sets of analyses were conducted where encoding trials (Hit and Miss trials) were further binned on the individual level by their even/odd trial numbers to establish the internal consistency of the subsequent memory paradigm. The internal consistency was assessed using ICC measures on the group-level for Hit rate, FA rate, and d'. RTs for subsequent Hit and subsequent Miss conditions were also assessed. Second, sets of analyses on 20 randomly drawn non-overlapping subsamples of 40 participants from the cross-sectional dataset were used to examine the consistency of Hit rate, FA rate, d', and RT across subsamples (using *t*-tests and Bayes Factors, https://klabhub.github.io/bayesFactor/). Last, sets of analyses were conducted on the test-retest data to assess differences in Hit rate, FA rate, d' as well as RTs between visits by calculating within-subject consistency (using ICC), and between-visit averages (using *t*-tests and Bayes Factors).

2.6. Reliability analysis of fMRI measures

2.6.1. Preprocessing and statistical analysis

Functional imaging data were analyzed with the SPM12 package (Wellcome Department of Imaging Neuroscience, London, UK). Images were motion-corrected, normalized to the Montreal Neurological Institute (MNI) template, and smoothed with an 8 mm full-width half-maximum Gaussian kernel. The same fMRI analysis pipeline was used for both the cross-sectional data and test-retest data.

A general linear model (GLM) was constructed for each participant visit. The GLM included regressors of interest based on subsequent memory outcomes (Hit or Miss) and scene complexity (high or low, Chai et al., 2010) for each of the three encoding runs. Because scene complexity was not a focus in this study, in all subsequent contrasts, regressors were combined across scene complexity. An additional regressor was included in the GLM for scenes with incorrect or no encoding responses (indoor/outdoor) to reduce possible confounds due to insufficient attention. Each encoding trial was modeled as an impulse function, convolved with a canonical model of the hemodynamic response function. Temporal derivatives were included for all conditions and were treated as regressors of no-interest. We controlled for motion in all participant visits in both the cross-sectional and the test-retest samples. For each run, 7 motion parameters were included in the model, and outlier volumes were controlled, by including covariates calculated through the Artifact Detection Tools (ART; http://www.nitrc.org/projects/artifact_detect/; an outlier is defined as global mean intensity > 3 SD or framewise displacement > 1 mm). To measure neural activation while performing the memory task and neural activation specifically related to memory formation, we computed 2 contrasts of interest for each individual:

(1) Encoding All Trials [(Subsequent Hit, 0.5 + Subsequent Miss, 0.5) vs. implicit baseline]
(2) Subsequent Memory Effect (Subsequent Hit, 1 vs. Subsequent Miss, −1)

Group-level analyses were conducted for each participant by combining individual-level contrast maps with a one-sample *t*-test. In order to fully appreciate the differences in the patterns between different group-level analyses, we visualized the results at a liberal threshold of $p < .05$, 100 contiguous voxels.

2.6.2. Reliability quantification of fMRI measures

We quantified the reliability of activations depending on assessing group or individual-level effects. For group-level fMRI results, we first generated two whole brain maps (e.g., *t*-maps for all participants in Subsample 1 and all participants in Subsample 2). An intra-voxel ICC can then be used to quantify the overall consistency by comparing the similarity between two *maps* (ICCv; Caceres et al., 2009; Raemaekers et al., 2007; Towgood et al., 2015). Effectively, in the ICCv calculation, we string all the voxels of each group-level map into one vector and quantify the consistency between the two vectors.

For individual-level reliability, we calculated one ICC value per each voxel. With each voxel, we obtained two values for all participants and visits, leading to a 24 (Participants) x 2 (Visits) matrix per voxel. This allows one ICC value to be calculated for each voxel. Then, by applying this approach to all voxels in the brain, we obtained a map of ICC values that visualized the level of reliability across all regions of the brain (as implemented in Caceres et al., 2009).

2.6.3. Cross-sectional sample internal paradigm consistency: split half by trial

To examine the reliability of the paradigm itself, we conducted a split-half trial analysis of the subsequent memory task. In the cross-sectional sample, we separated all encoding trials into two halves (even and odd) and compared the behavioral measures and fMRI effects between the group results of all participants. For behavioral measures, we reported the mean and SD statistics of age and memory performance, separately for even and odd trials. For fMRI results, in order to cleanly separate the neural activations for even and odd trials, we conducted the beta-series analysis on each participant (Mumford et al., 2012; Rissman et al., 2004). In this analysis, we constructed one GLM for each encoding trial, where the trial was modeled separately from all other regressors, and a beta value was generated for that trial (the Least Squares – Separate (LSS) method, as adapted from https://github.com/tsalo/misc-fmri-code/tree/master/lss). This LSS method was then repeated for all trials. To reduce the effect of motion in the beta-series analysis, we excluded trials that were 2 volumes before or 5 volumes after any motion spike as identified by ART (Power et al., 2012). One participant was excluded from the beta-series analysis due to insufficient number of trials after motion censoring specific to the LSS method. After the beta-series analysis, we split all trials into two halves for each participant. We combined individual-level results with one-sample *t*-tests separately for even and for odd trials, resulting in Encoding Even and Encoding Odd contrasts (compared to implicit baseline). We then conducted two-sample *t*-tests comparing Hit and Miss conditions separately for even and odd trials, resulting in SME Even and SME Odd contrasts. To generate group-level results, we conducted one-sample *t*-tests to combine two individual-level contrasts for even and odd trials separately, resulting in 4 group-level maps (Encoding Even, Encoding Odd, SME Even, and SME Odd). We quantified the consistency (ICC) in the group-level results between the even/odd splits for both behavioral and fMRI results.

2.6.4. Cross-sectional sample between-subject reliability: group-level subsamples

To examine the reliability in the results of subsequent memory paradigm across different participant selections, we conducted 20 random draws of subsamples, each consisting 2 non-overlapping groups of 40 participants. Within each draw, we compared group-level results for both behavioral and fMRI measures. For behavioral measures, we reported the statistics of age and memory performance for each subsample. For fMRI measures, we computed group-level results combining individual-level contrasts (Encoding All Trials and SME) for each subsample. We quantified the level of consistency in the group-level results between the two subsamples for both behavioral and fMRI results.

2.6.5. Test-retest sample reliability: examined at the group and individual level

Test-retest reliability was assessed by the correspondence in the activation maps on the group level and on the individual level. For the group-level reliability, we first calculated group effects separately for each visit from individual contrasts, both Encoding All Trials and SME. Group-level reliability was then computed to compare the group maps between two visits using ICCv. On the other hand, individual-level reliability was assessed by calculating one ICC value per voxel and repeating it across the whole brain. We utilized previously validated functions from the ICC toolbox to generate individual-level ICC maps (Caceres et al., 2009). To visualize individual-level reliability by brain region, we highlighted in the ICC maps regions that demonstrated at least fair reliability (ICC $>$ .4).

2.6.6. Reliability by brain region and fMRI contrast

To examine potential differences in the consistency and reliability of different brain regions, we assessed between-subject and test-retest reliability for four regions of interest (ROIs), including inferior frontal gyrus (IFG), superior frontal gyrus (SFG), hippocampus, and parahippocampal gyrus (PHG). We used bilateral structural ROIs generated from the SPM Anatomy Toolbox (https://www.fz-juelich.de/inm/inm-1/DE/Forschung/_docs/SPMAnatomyToolbox/SPMAnatomy-Toolbox_node.html). Similar to the whole brain analysis, between-subject and test-retest reliability (group- and individual-level) were calculated for each ROI.

3. Results

3.1. Reliability of behavioral measures

3.1.1. Cross-sectional sample internal paradigm consistency: split-half by trial

Between the two halves of trials, we observed comparable memory performance, as measured by d' of Even and Odd trials (d' of Even trials: 0.97 ± 0.54, d' of Odd trials: 0.95 ± 0.56; RT of even subsequent Hit trials: 1.08 ± 0.31, RT of odd subsequent Hit trials: 1.08 ± 0.32, Table 1, Fig. 1B). We observed excellent split-half reliability for memory performance (ICC for d' = 0.89) and excellent split-half reliability for RTs of subsequent Hit and subsequent Miss trials (ICCs for d' $>$ .90).

3.1.2. Cross-sectional sample between-subject reliability: group-level subsamples

There was no significant difference in the behavioral measures between the subsamples, for Hit rate (Subsample 1: .57 ± .15, Subsample 2: .59 ± .13; $p = .68$), FA rate (Subsample 1: .25 ± .14, Subsample 2: .27 ± .14; $p = .55$), or d' (Subsample 1: 0.97 ± 0.55, Subsample 2: 0.92 ± 0.54; $p = .67$), suggesting a good between-subject reliability in memory performance (see Table 1, Fig. 1C). Correcting for multiple comparisons (alpha = .05/5 = .01), there was a trend-level non-significant difference in RTs of subsequent Hit (Subsample 1: 1.16 s ± 0.33, Subsample 2: 1.00 s ± 0.29; $p = .03$) and subsequent Miss trials (Subsample 1: 1.13 s ± 0.30, Subsample 2: 1.00 s ± 0.28; $p = .07$). Bayes Factors did not provide evidence of differences in any of the comparisons. Overall, we observed high level of consistency in memory performance between two randomly selected subsamples.

3.1.3. Test-retest sample reliability: examined on the group and individual level

Memory performance calculated for each of the visits in the test-retest reliability sample was similar (Visit 1: 0.84 ± 0.42, Visit 2: 0.88 ± 0.52; $p = .62$; Table 1, Fig. 1D). In examining individual-level reliability in behavioral measures between two visits, we observed overall good test-retest reliability in memory performance (ICC for d' = .62, Fig. 1D, greyscale lines link individual participant performance across the two visits). We observed good test-retest reliability in RTs of subsequent Hit (ICC for RTs = .66) and subsequent Miss trials (ICC for RTs = .62) (Table 1).

3.1.4. Number of trials for reliability analysis

Prior to conducting fMRI reliability analyses, we assessed the number of trials per conditions in the data in each sample. We verified comparable number of trials for Hit and Miss trials in all analyses. Specifically, we obtained data in the cross-sectional sample from 66.75 ± 16.46 Hit trials (range: 16–104) and 48.16 ± 16.43 Miss trials (range: 15–99). Similarly, in the test-retest sample we obtained data from 69.06 ± 11.07 Hit trials (range: 45–97), and 47.85 ± 10.80 Miss trials (range: 21–74). The number of Hit and Miss trials differed by age in the cross-sectional sample (Hit: $r(83) = .34$, $p < .001$, Miss: $r(83) = -.25$, $p = .02$), but not in the test-retest sample (all $ps > .05$).

Table 1
Reliability of behavioral measures.

	Age	Hit rate	FA rate	d'	Hit RT	Miss RT
Cross-sectional sample	16.51 ± 4.73	.58 ± .14	.25 ± .14	0.96 ± 0.54	1.07 ± 0.31	1.07 ± 0.29
Even trials		.59 ± .14	.25 ± .14	0.97 ± 0.54	1.08 ± 0.31	1.06 ± 0.29
Odd trials		.58 ± .15	.26 ± .15	0.95 ± 0.56	1.08 ± 0.32	1.07 ± 0.30
Paradigm Consistency (ICC)		.82	.80	.89	.95	.90
Subsample 1	16.08 ± 4.44	.57 ± .15	.25 ± .14	0.97 ± 0.55	1.16 ± 0.33	1.13 ± 0.30
Subsample 2	16.46 ± 4.89	.59 ± .13	.27 ± .14	0.92 ± 0.54	1.00 ± 0.29	1.00 ± 0.28
Subsample comparison	$p = .72$	$p = .68$	$p = .55$	$p = .67$	$p = .03$	$p = .07$
Bayes factor	0.25	0.25	0.27	0.25	1.79	1.02
Test-retest sample						
Visit 1	13.27 ± 3.14	.58 ± .09	.27 ± .10	0.84 ± 0.42	1.05 ± 0.19	1.05 ± 0.19
Visit 2	13.35 ± 3.14	.60 ± .09	.28 ± .13	0.88 ± 0.52	1.14 ± 0.27	1.18 ± 0.29
Test-retest Reliability (ICC)		.26	.59	.62	.66	.62
Visit comparison		$p = .46$	$p = .78$	$p = .76$	$p = .18$	$p = .06$
Bayes factor		0.36	0.30	0.30	0.61	1.29

FA: false alarm. d' = Z(Hit rate) – Z(Miss rate). RT: reaction time. Bayes Factors defined as BF10.

3.2. Reliability of fMRI measures

3.2.1. Cross-sectional sample internal paradigm consistency: split-half by trial

We next assessed the reliability of fMRI measures in the subsequent memory paradigm. Given the potential influence from the choice of fMRI contrast on reliability measures, we examined two contrasts of interest: (1) Encoding All Trials (all trials regardless of subsequent memory outcome) and (2) Subsequent Memory Effects (SME, subsequent Hit > subsequent Miss trials). We calculated group-level contrast maps separately for the two contrasts of interest and compared the group-level results between even and odd trials. For the Encoding All Trials contrast, we observed typical patterns of activation and deactivation commonly associated with performing a cognitive task. Specifically, task-related activation was observed in inferior frontal gyrus, superior parietal lobe, and lateral occipital lobe; task-related deactivation was observed in regions that are part of the Default Mode Network (DMN), including medial PFC (mPFC), precuneus, and lateral parietal lobe (Fig. 2A). Comparing group-level fMRI patterns between even and odd trials, we found excellent reliability between the maps generated with each half of the trials (ICCv = .89). For the SME contrast, we observed memory-related activation in inferior frontal gyrus, superior parietal lobe, parahippocampal gyrus, and lateral occipital lobe; we observed memory-related deactivation in superior frontal gyrus, mPFC, precuneus, and lateral parietal lobe. Comparing group-level fMRI patterns between even and odd trials, we found modest

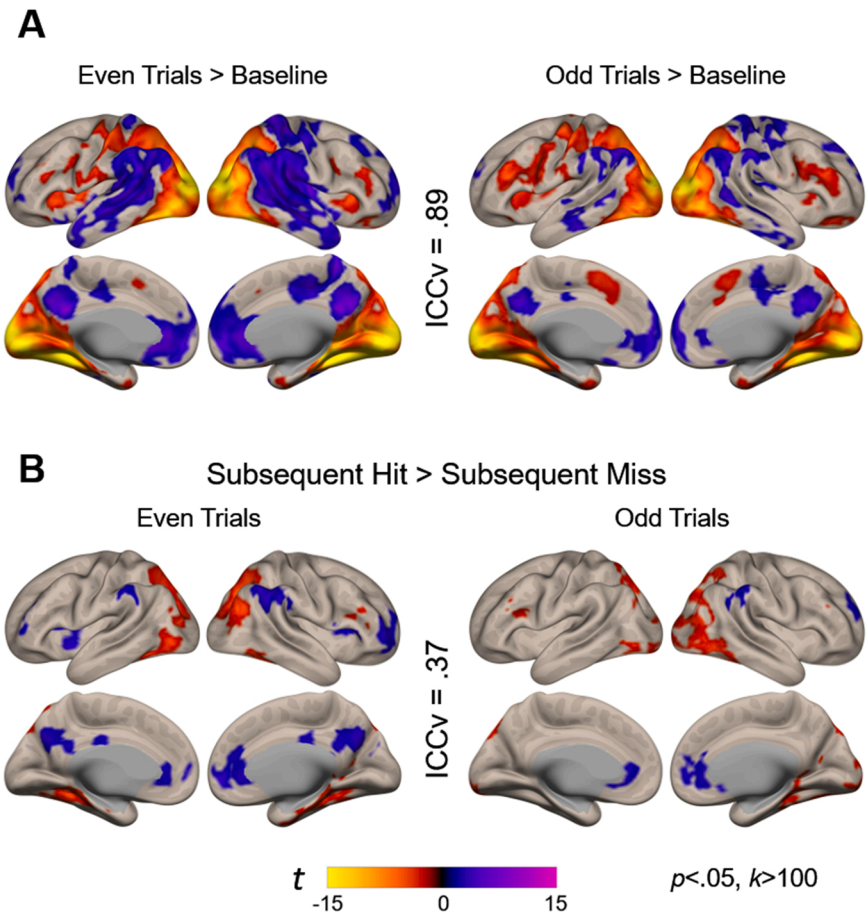


Fig. 2. Reliability of the experimental paradigm. Reliability was assessed by comparing splits based on the even (left) or odd (right) numbered trials during encoding. A. Excellent reliability was observed in the group-level fMRI activation maps for even (A, left) and odd (A, right) trials (ICCv = .89), without accounting for the subsequent memory outcome of each trial (Encoding All Trials). B. Modest reliability was observed in the group-level fMRI results when assessing the subsequent memory effects (SME; Hit > Miss) comparing split-halves based on the even (B, left) or odd (B, right) numbered trials during encoding (ICCv = .37). We observed positive SME in frontal, temporal, and occipital cortex, and negative SME in parietal and medial prefrontal cortex. Activations maps were overlaid on a surface mesh.

split-half reliability (ICCv = .37) (Fig. 2B).

3.2.2. Cross-sectional sample between-subject reliability: group-level subsamples

Next, to understand the consistency of fMRI results across different subsamples of participants, we examined the between-subject group-level reliability of the activation maps generated based on the two contrasts of interest. From the cross-sectional sample, between-subject reliability measures were assessed with a total of 20 random draws of subsamples, where we selected 2 non-overlapping groups of 40 participants. While we did not intentionally match subsamples for age to reduce potential bias, across all 20 draws, subsamples did not differ by age (all *ps* > .17). We calculated the ICCv for all 20 draws and for both contrasts of interest. Overall, the reliability measurement was unaffected by specific subsample selections (ICCv for Encoding All Trials contrast: .95 ± .01; ICCv for SME contrast: .75 ± .08, See Supplementary Fig. 3 for the distribution of ICC values for each ROI across the 20 draws). Therefore, we present one of the random draws comparing non-overlapping subsamples below for ease of interpretation (Fig. 3; $p < .05$, 100 contiguous voxels for visualization purposes).

For the Encoding All Trials contrast, we observed task-related activation in inferior frontal gyrus, superior parietal lobe, hippocampus, parahippocampal gyrus, and lateral occipital lobe. We observed task-related deactivation in the DMN, including mPFC, precuneus, and lateral parietal lobe (Fig. 3A). Comparing group-level fMRI patterns between two subsamples, we found excellent between-subject reliability (ICCv = .94) for activations encoding indoor and outdoor scenes.

For the SME contrast, we observed positive SME in inferior frontal gyrus, superior parietal lobe, hippocampus, parahippocampal gyrus, and lateral occipital lobe. We observed negative SME in superior frontal gyrus, mPFC, precuneus, and lateral parietal lobe (Fig. 3B). Comparing group-level contrasts between two subsamples, we found good between-subject consistency (ICCv = .73) for subsequent memory effects.

3.2.3. Test-retest sample reliability

3.2.3.1. Test-retest sample reliability examined on the group level. We further determined the reliability of fMRI activation maps in the test-retest sample (participants scanned twice with the two visits spaced one month apart). We first examined group-level data by contrasting activation maps generated based on data from Visit 1 to those generated from Visit 2. Overall, we observed very similar group-level activation maps across the two visits, indicating high group-level consistency (Fig. 4; $p < .05$, 100 contiguous voxels for visualization purposes).

For the Encoding All Trials contrast, we observed task-related activation in inferior frontal gyrus, superior parietal lobe, hippocampus, parahippocampal gyrus, and lateral occipital lobe. We observed task-related deactivation in several regions of the DMN, including mPFC, inferior parietal lobe, and superior frontal gyrus (Fig. 4A). Comparing group-level fMRI patterns between two visits, we found excellent test-retest reliability (ICCv = .91) for activations encoding indoor and outdoor scenes.

For the SME contrast, we observed positive SMEs in inferior frontal gyrus, hippocampus, parahippocampal gyrus, and lateral occipital lobe. We observed negative SMEs in superior frontal gyrus, mPFC, precuneus, and lateral parietal lobe (Fig. 4B). Comparing group-level contrasts between two visits, we found good between-subject consistency (ICCv = .70) for subsequent memory effects.

To examine potential age effects on the group-level reliability across the two visits, we separated the dataset by the median age of the current sample (13 years) and calculated the ICC maps of both contrasts

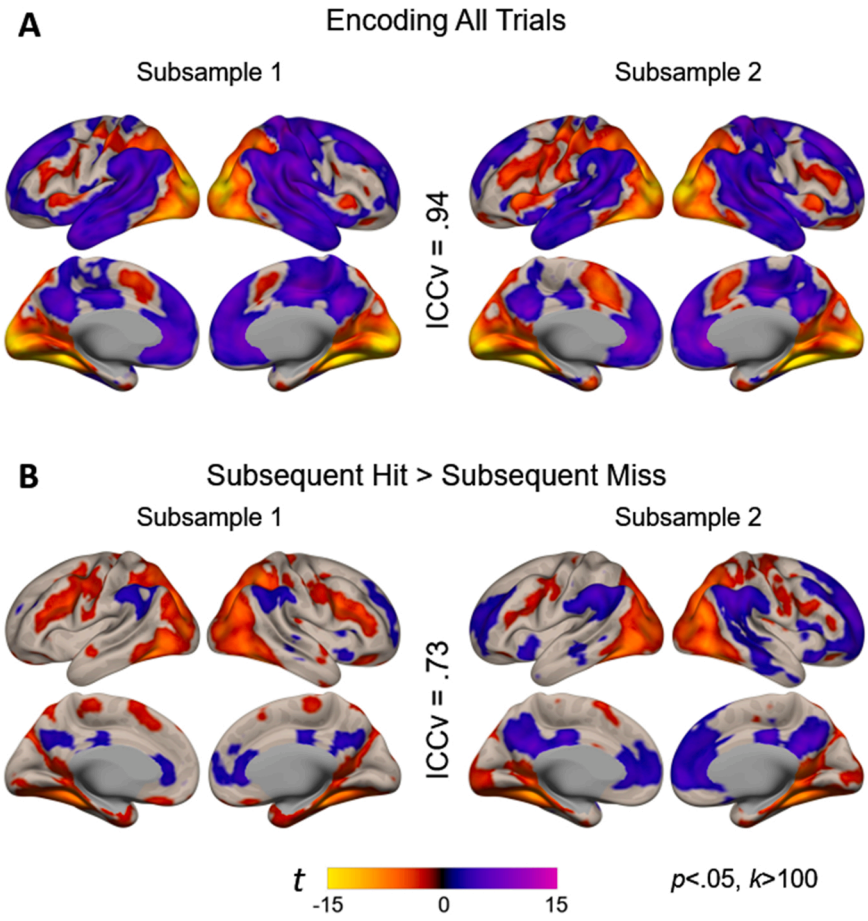


Fig. 3. Reliability between different sampling of participants. Between-subject reliability was assessed by comparing the fMRI results from two non-overlapping subsamples (40 participants each) from the cross-sectional dataset. A. For the Encoding All Trials contrast (all trials vs. implicit baseline), excellent reliability was observed for group-level fMRI results (ICCv = .94). B. For the Subsequent Memory contrast (SME; Hit > Miss), good group-level reliability was observed (ICCv = .73).

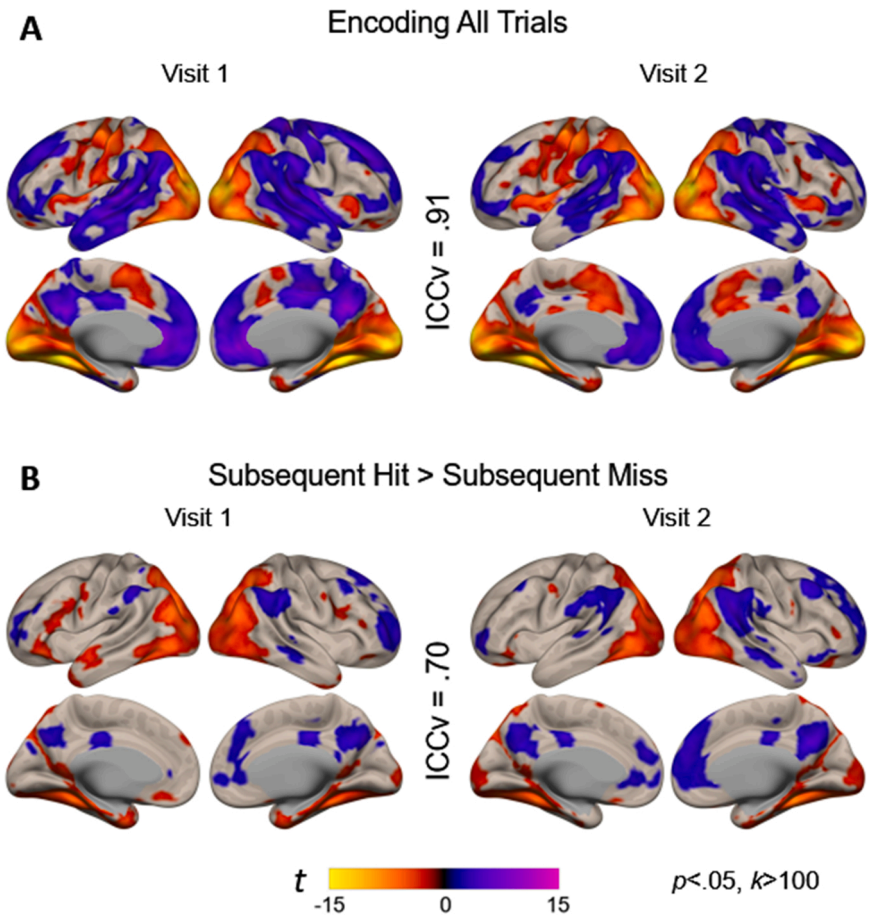


Fig. 4. Group-level test-retest reliability by contrast. A. For the Encoding All Trials contrast (all trials vs. implicit baseline), excellent reliability was observed for group-level fMRI results between two visits (ICCv = .91). B. For the Subsequent Memory contrast (SME; Hit > Miss), good group-level reliability was observed (ICCv = .70).

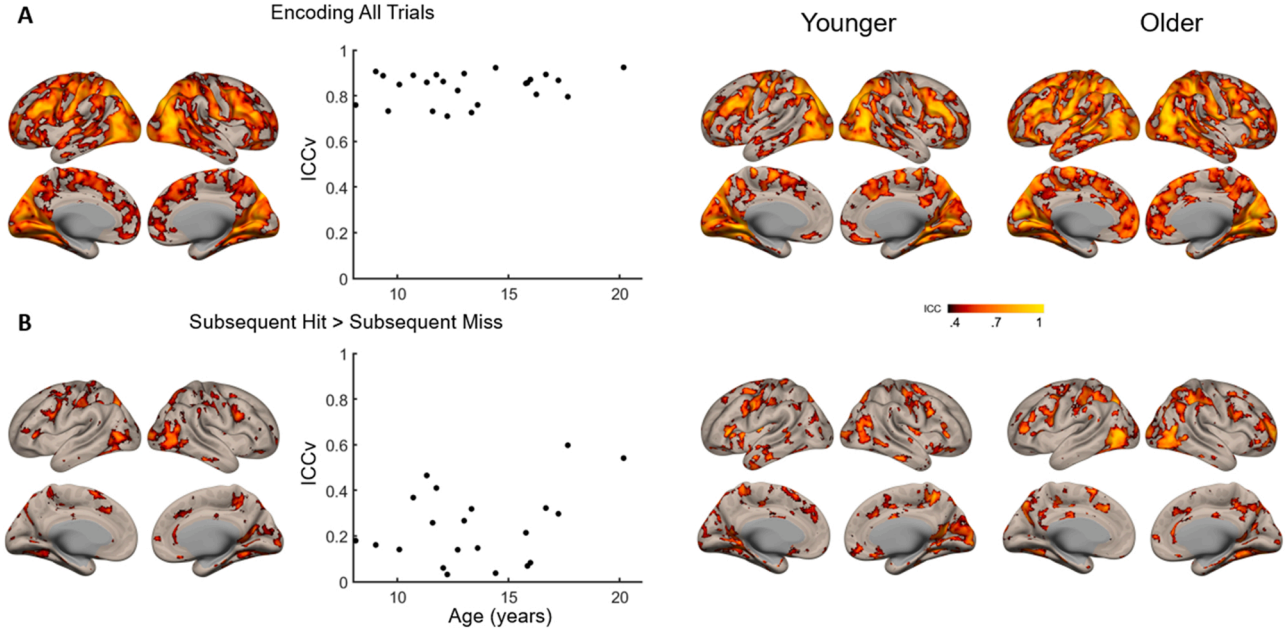


Fig. 5. Test-retest reliability as measured by ICC across the brain for the Encoding All Trials (A) and SME contrast (B). For both contrasts, good test-retest reliability was observed in bilateral inferior frontal gyrus, right parahippocampal gyrus, and bilateral lateral occipital lobe (ICC > .6). Hippocampus showed poor reliability in the SME contrast (ICC < .4). Individual intravoxel reliability (ICCv) did not correlate with age for the Encoding All Trials contrast ($r(22) = .02$, $p = .93$, top middle) or the SME contrast ($r(22) = .35$, $p = .09$, bottom middle).

separately for younger (age: 10.71 ± 1.46, [8.12, 12.71], 6 M:6 F, Supplementary Fig. 1) and older (age: 15.83 ± 2.05, [13.02, 20.18], 6 M:6 F, Supplementary Fig. 2) participants. For both younger and older participants, we found good reliability in IFG, PHG, and middle occipital lobe for Encoding All Trials and SME contrasts, similar to the findings based on the full sample. These findings suggest minimal age differences when assessing test-retest reliability at the group level.

3.2.3.2. Test-retest sample reliability examined on the individual level. Next, we investigated test-retest reliability on the individual level, first of the Encoding All Trials and then of the SME contrast. We first generated whole-brain reliability maps by calculating per-voxel ICC values between two visits and then thresholded reliability maps by ICC > .4 (Fig. 5). For the Encoding All Trials contrast, good reliability was found in several cortical regions, including bilateral IFG, PHG, lateral occipital lobe, cuneus, and posterior regions of the hippocampus (ICC > .6). In contrast, poor test-retest reliability was found in subcortical regions (ICC < .4; Fig. 5A). For the SME contrast, good reliability was observed in bilateral IFG, right PHG, and bilateral lateral occipital lobe (ICC > .6), whereas poor reliability was observed in the hippocampus (ICC < .4; Fig. 5B).

As we were interested in age effects of reliability in the fMRI activations, for both Encoding All Trials and SME contrasts, we computed an ICCv per individual and correlated the ICCv values with age. We found no age effects in the individual-level reliability for the Encoding All Trials contrast ($r(22) = .02$, $p = .93$) and a non-significant trending age effect for the SME contrast ($r(22) = .35$, $p = .09$; Fig. 5).

3.3. Reliability by brain region and fMRI contrast

In order to examine the potential differences in the reliability of fMRI activations in different brain regions and for different contrasts, we selected four bilateral ROIs that are commonly known to elicit SME: IFG, SFG, hippocampus, and PHG, as defined in the Anatomy toolbox (for details, see Methods). We quantified the level of between-subject reliability and test-retest reliability for each ROI.

Consistent with the whole-brain analyses, we found good to excellent reliability in group-level ROI effects for both cross-sectional and test-retest samples (Table 2). Reliability was higher for Encoding All Trials contrast than SME contrast. Notably, the test-retest reliability for SME was poor to fair across multiple ROIs on the individual level, with the hippocampus showing comparatively lower reliability. Overall, while group-level fMRI results based on the subsequent memory paradigm are highly consistent across different sample selections and across time, individual-level results were less reliable. Reliability was higher for the Encoding All Trials contrast than the SME contrast. To help appreciate the range and variance of fMRI activations in these ROIs between the contrasts, we included the histograms in the Supplementary Material (Supplementary Fig. 4).

Table 2
Reliability by region and contrast.

	IFG	SFG	Hippocampus	PHG
Between-Subject Reliability (Cross-sectional)				
Encoding All Trials	.67	.80	.94	.95
Subsequent Memory	.74	.61	.47	.79
Test-retest Reliability (Group-level)				
Encoding All Trials	.74	.73	.87	.90
Subsequent Memory	.51	.72	.45	.46
Test-retest Reliability (Individual-level)				
Encoding All Trials	.53	.46	.49	.72
Subsequent Memory	.23	.19	.10	.16

Between-subject and test-retest reliability as measured by ICC (or ICCv) values. IFG: inferior frontal gyrus; SFG: superior frontal gyrus; PHG: parahippocampal gyrus.

4. Discussion

In this study, we investigated the reliability of behavioral and fMRI measures in children and young adults while they performed a commonly used subsequent memory task. In both a cross-sectional and a test-retest sample, we found excellent group-level reliability for participants undertaking the subsequent memory paradigm, for both the encoding task and subsequent memory effects. However, the consistency of activation patterns within individuals was modest, especially for the subsequent memory contrast. Of particular importance to memory development, we observed consistent activation patterns in frontal, parietal, and occipital cortices, but comparatively lower test-retest reliability in subcortical regions and the hippocampus.

Several important aspects of consistency and reliability were assessed in this study. First, we set to establish the internal consistency of the task, the subsequent memory paradigm, in generating reliable behavioral measures of memory in a developmental sample. By repeating the analyses after binning trials based on their numbers, we observed excellent consistency in the effects comparing even to odd trials in the behavior performance. These results provide initial evidence for the internal consistency of the task. Second, by splitting the dataset into two subsamples, we observed a comparable level of memory performance between subsamples. Moreover, comparing memory performance within the same participants who completed the task twice as part of the test-retest study, we observed good within-subject consistency in memory performance and reaction time. The high degree of reliability in the behavioral data of our developmental samples is in line with other studies on young adults and adolescence investigating the reliability for behavior measures in and out of the scanner (Hedge et al., 2018; Van Den Bulk et al., 2013), and provides the foundation to assess the reliability of fMRI measures.

Next, we found overall good to excellent group-level reliability in both contrasts of interest, Encoding All Trials and SME. These findings are consistent with previous studies showing high reliability in group-level activation across different experimental paradigms with fMRI (Aron et al., 2006; Plichta et al., 2012; Raemaekers et al., 2007). We can therefore infer that, when we average the functional activation for an fMRI paradigm with a group of participants, we can reliably establish the activation pattern for this paradigm. Combining results from a reasonably large sample, fMRI has been shown to reliably identify memory-related regions, where consistent activations across the PFC and MTL regions have been observed (Kim, 2011; Spaniol et al., 2009). Our findings confirmed high reliability in group-level results that is unaffected by different sampling and extended these findings to developmental samples.

When we investigated reliability in the test-retest dataset, we observed higher group-level reliability compared to individual-level reliability. Specifically, while group-level averaging of more than 20 participants produced highly consistent results, individual results of the same participant tested twice were much less consistent. We speculate that the fMRI scans, due to multiple sources of noise (e.g., susceptibility and motion), show large differences on the individual-level, resulting in low reliability between visits. However, by conducting the group-level analysis, the random noise in individuals is canceled out, leading to more reliable group-level effects. Overall, we show that in a developmental sample, the subsequent memory paradigm can produce reliable group-level results, but less reliable individual-level results.

In addition, the level of reliability differs by the choice of contrast and region of interest. The choice of contrast had a large effect on the estimation of reliability. Curiously, fMRI contrasts that are less specific (e.g., Encoding All Trials vs. implicit baseline) showed higher reliability than contrasts that are more specific (e.g., SME). We reason that the subtraction process in generating an effective contrast removed individual idiosyncrasies that contribute to high reliability (See also Infantolino et al., 2018). Our results thus highlight the paradox between reliability and specificity: when targeting a very specific cognitive

construct, reliability suffers. Our findings also demonstrated regional differences in the level of reliability. While good reliability was observed across most regions for the Encoding All Trials contrast, poor reliability was observed for the SME contrast, especially in the hippocampus, suggesting that memory-related activation in the hippocampus may be unreliable on an individual level. Previous studies investigating the reliability of memory-related fMRI activation generally found low reliability in the hippocampus (Brandt et al., 2013; Clément and Belleville, 2009, but see Putcha et al., 2011). Our findings contribute to the accumulating evidence showing modest test-retest reliability in widely validated tasks (Bennett and Miller, 2013, 2010; Caceres et al., 2009; Elliott et al., 2020). The low within-individual reliability of memory effects observed in this study suggests that assessment of reliability should be incorporated for adequate interpretation of memory-related effects.

Although we hypothesized age differences in test-retest reliability with fMRI based on one a previous study that reported higher reliability in adults compared to children (Koolschijn et al., 2011), we observed no such effect in this sample. Several possible explanations may account for the differences between studies. First, different paradigms were used in the two studies – in this study, a subsequent memory paradigm was used, and in the other, a performance monitoring task with significant motor components. It is therefore possible that age effects are limited when assessing the reliability of SME. Second, in this study, we carefully controlled for factors like motion, which is known to generate spurious effects in developmental studies. Controlling for motion may have removed possible spurious age effects in reliability estimates. Third, we kept a relatively short time gap between two visits and assessed reliability that is not compounded by possible developmental changes. This contrasts with the prior report that included a 3.5-year gap between visits (Koolschijn et al., 2011). Together, our findings suggest it is possible that there are no systematic age differences in reliability when identifying memory-related activation with fMRI. However, low level of reliability, regardless of whether it differs by age or not, limits our ability to investigate individual differences above and beyond measurement error.

As expected with developmental samples, we do find age-related differences in memory performance and number of Hit trials, with high performance and greater number of Hit trials in older participants. A potential limitation of this study is the influence of age differences in memory performance and number of trials on the reliability. We argue that, while these performance related behavioral factors may affect SME, these effects are minimized by the fact that we included more than 15 trials for each condition to allow for sufficient power when conducting the fMRI analysis. Furthermore, despite the possible difference in age-related SME effects, we did not find a significant age effect in the reliability of the fMRI response, suggesting that the memory performance and different number of trials may not be a prominent factor that influences reliability.

5. Future directions

In this study, we found excellent group-level reliability with the subsequent memory paradigm, but reduced individual-level reliability for the specific contrast targeting the memory construct (subsequently remembered vs. forgotten). Relatively low reliability was found in the hippocampus. Based on our findings in the current study and the other reliability studies reviewed above, we make several recommendations for future developmental research:

First, given the importance of reliability in grounding the findings in developmental studies, it is ideal to incorporate a reliability assessment of selected research paradigm. If resources are available, a test and retest approach can be embedded in the study design to examine both group- and individual-level reliability. On the other hand, if only a cross-sectional design is possible, split-half reliability or reliability between subsamples can be examined.

Second, increase the scan length whenever possible. Our current analysis utilized a common memory encoding task with 120 trials. We had 3 functional runs totaling 12 min, a typical task length for fMRI. While reliability has shown to be adequate to address questions on a group-level results, e.g., to identify shared brain regions related to memory formation, the task length may be inadequate to examine individual-level differences or developmental trajectory longitudinally. While it is widely known that fMRI signal is intrinsically noisy, researchers have historically preferred short fMRI task that are shown to activate target brain regions across all participants so that more tasks can be packed into one scan session. This practice may have placed a limit on the reliability these fMRI tasks can achieve. Based on recent studies showing that increasing scan time to 27 min can improve test-retest reliability of resting-state functional connectivity to $r > .8$ (Gordon et al., 2017), the "cure" for low task-based fMRI reliability might be an obvious one: scan more.

Consistent with the idea of obtaining richer data, researchers in recent years have considered two approaches: increasing either the breadth or the depth of the data. On one hand, researchers can increase the number of participants in the study (the big data approach); on the other, researchers can collect large amount of data from each participant (the high precision approach). Consistent with the big data approach, it has been shown that increasing sample size beyond 100 will continue to increase reliability, although the degree of reliability is heavily influenced by the specific task and contrast selected, as we and others have also shown (Bossier et al., 2020; Turner et al., 2018). Public data sharing initiatives such as Human Connectome Project (HCP) and the Adolescent Brain Cognitive Development (ABCD) Study are instrumental in improving the reproducibility of fMRI data. Regarding the high precision approach, recent studies using Midnight Scan Club data (Gordon et al., 2017) have shown that cortical functional connectivity was highly reliable within an individual, given ample amount of data per individual. Combining the big data and high precision approaches, leading studies in developmental science have been able to detect subtle differences in individual functional topography that evolves with youth in a data of 693 participants (Cui et al., 2020). With the combined approach, researchers may be able to unblur the developmental picture by minimizing the noise in fMRI data.

6. Conclusions

In sum, we investigated the reliability of fMRI activations in a subsequent memory paradigm using a cross-sectional and a test-retest sample. We found excellent reliability with the subsequent memory paradigm for group-level contrasts related to general memory encoding (all encoding trials vs. implicit baseline), but reduced reliability for individual-level results, especially for the contrast targeting the memory construct (remembered vs. forgotten). In addition, we identified regional specificity in the degree of reliability. Reliability was good to excellent in frontal, parietal, and occipital cortices, but poor in subcortical regions and the hippocampus. These findings highlight potential limitations using task-based fMRI to understand development. In the future, by leveraging bigger and denser data and incorporating reliability analyses routinely in developmental studies, we may better ensure the reproducibility of our findings and safeguard the accumulation of knowledge in developmental sciences.

Funding

The work was supported by the National Institutes of Health [grant number R01MH107512 (NO)].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence

the work reported in this paper.

Data availability

The data that support the findings of this study are openly available in NIMH Data Archive (NDA) at https://nda.nih.gov/ (DOI 10.15154/1524262). Additional data related to the paper may be requested from the authors.

Acknowledgments

We thank Bryn Thompson, Dana McCall, Sruthi Ramesh, David Zhijian Chen, and Pavan Jella Kumar for data collection, John France, Raymond Viviano for the help with data analyses. We thank Cindy Lustig, Vaibhav Diwadkar, Ana M. Daugherty, and Naftali Raz for insightful discussions. We thank baby Luo Fei for giving all of us joy.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.dcn.2021.101037.

References

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. https://doi.org/10.1016/j.neuroimage.2005.08.010.

Bennett, C.M., Miller, M.B., 2013. FMRI reliability: influences of task and experimental design. Cogn. Affect. Behav. Neurosci. 13, 690–702. https://doi.org/10.3758/s13415-013-0195-1.

Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging? Ann. N. Y. Acad. Sci. 1191, 133–155. https://doi.org/10.1111/j.1749-6632.2010.05446.x.

Boenniger, M.M., Diers, K., Herholz, S.C., Shahid, M., Stöcker, T., Breteler, M.M.B., Huijbers, W., 2021. A functional MRI paradigm for efficient mapping of memory encoding across sensory conditions. Front. Hum. Neurosci. 14, 1–16. https://doi.org/10.3389/fnhum.2020.591721.

Bossier, H., Roels, S.P., Seurinck, R., Banaschewski, T., Barker, G.J., Bokde, A.L.W., Quinlan, E.B., Desrivières, S., Flor, H., Grigis, A., Garavan, H., Gowland, P., Heinz, A., Ittermann, B., Martinot, J.L., Artiges, E., Nees, F., Orfanos, D.P., Poustka, L., Fröhner Dipl-Psych, J.H., Smolka, M.N., Walter, H., Whelan, R., Schumann, G., Moerkerke, B., 2020. The empirical replicability of task-based fMRI as a function of sample size. Neuroimage 212, 1–12. https://doi.org/10.1016/j.neuroimage.2020.116601.

Brandt, D.J., Sommer, J., Krach, S., Bedenbender, J., Kircher, T., Paulus, F.M., Jansen, A., 2013. Test-retest reliability of fMRI brain activity during memory encoding. Front. Psychiatry 4, 1–9. https://doi.org/10.3389/fpsyt.2013.00163.

Van Den Bulk, B.G., Koolschijn, P.C.M.P., Meens, P.H.F., Van Lang, N.D.J., Van Der Wee, N.J.A., Rombouts, S.A.R.B., Vermeiren, R.R.J.M., Crone, E.A., 2013. How stable is activation in the amygdala and prefrontal cortex in adolescence? A study of emotional face processing across three measurements. Dev. Cogn. Neurosci. 4, 65–76. https://doi.org/10.1016/j.dcn.2012.09.005.

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. https://doi.org/10.1016/j.neuroimage.2008.12.035.

Chai, X.J., Ofen, N., Jacobs, L.F., Gabrieli, J.D.E., 2010. Scene complexity: influence on perception, memory, and development in the medial temporal lobe. Front. Hum. Neurosci. 4, 21. https://doi.org/10.3389/fnhum.2010.00021.

de Chastelaine, M., Wang, T.H., Minton, B., Muftuler, L.T., Rugg, M.D., 2011. The effects of age, memory performance, and callosal integrity on the neural correlates of successful associative encoding. Cereb. Cortex 21, 2166–2176. https://doi.org/10.1093/cercor/bhq294.

Cicchetti, D., 2001. Methodological commentary the precision of reliability and validity estimates re-visited: distinguishing between clinical and statistical significance of sample size requirements. J. Clin. Exp. Neuropsychol. 23, 695–700. https://doi.org/10.1076/jcen.23.5.695.1249.

Clément, 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. https://doi.org/10.1002/hbm.20827.

Cui, Z., Li, H., Xia, C.H., Larsen, B., Adebimpe, A., Baum, G.L., Cieslak, M., Gur, R.E., Gur, R.C., Moore, T.M., Oathes, D.J., Alexander-Bloch, A.F., Raznahan, A., Roalf, D. R., Shinohara, R.T., Wolf, D.H., Davatzikos, C., Bassett, D.S., Fair, D.A., Fan, Y., Satterthwaite, T.D., 2020. Individual variation in functional topography of association networks in youth, 340-353.e8 Neuron 106. https://doi.org/10.1016/j.neuron.2020.01.029.

DeMaster, D., Pathman, T., Ghetti, S., 2013. Development of memory for spatial context: hippocampal and cortical contributions. Neuropsychologia 51, 2415–2426. https://doi.org/10.1016/j.neuropsychologia.2013.05.026.

Elliott, M.L., Knodt, A.R., Ireland, D., Morris, M.L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T.E., Caspi, A., Hariri, A.R., 2020. What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. Psychol. Sci. 31, 792–806. https://doi.org/10.1177/0956797620916786.

Fröhner, J.H., Teckentrup, V., Smolka, M.N., Kroemer, N.B., 2019. Addressing the reliability fallacy in fMRI: similar group effects may arise from unreliable individual effects. Neuroimage 195, 174–189. https://doi.org/10.1016/j.neuroimage.2019.03.053.

Ghetti, S., Bunge, S.A., 2012. Neural changes underlying the development of episodic memory during middle childhood. Dev. Cogn. Neurosci. 2, 381–395. https://doi.org/10.1016/j.dcn.2012.05.002.

Ghetti, S., DeMaster, D.M., Yonelinas, A.P., Bunge, S. a, 2010. Developmental differences in medial temporal lobe function during memory encoding. J. Neurosci. 30, 9548–9556. https://doi.org/10.1523/JNEUROSCI.3500-09.2010.

Gordon, E.M., Laumann, T.O., Gilmore, A.W., Newbold, D.J., Greene, D.J., Berg, J.J., Ortega, M., Hoyt-Drazen, C., Gratton, C., Sun, H., Hampton, J.M., Coalson, R.S., Nguyen, A.L., McDermott, K.B., Shimony, J.S., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., Nelson, S.M., Dosenbach, N.U.F., 2017. Precision functional mapping of individual human brains, 791-807.e7 Neuron 95. https://doi.org/10.1016/j.neuron.2017.07.011.

Gorgolewski, K.J., Storkey, A., Bastin, M.E., Whittle, I.R., Wardlaw, J.M., Pernet, C.R., 2013. A test-retest fMRI dataset for motor, language and spatial attention functions. Gigascience 2, 6–9. https://doi.org/10.1186/2047-217X-2-6.

Güler, O.E., Thomas, K.M., 2013. Developmental differences in the neural correlates of relational encoding and recall in children: an event-related fMRI study. Dev. Cogn. Neurosci. 3, 106–116. https://doi.org/10.1016/j.dcn.2012.07.001.

Harrington, G.S., Tomaszewski Farias, S., Buonocore, M.H., Yonelinas, A.P., 2006. The intersubject and intrasubject reproducibility of FMRI activation during three encoding tasks: implications for clinical applications. Neuroradiology 48, 495–505. https://doi.org/10.1007/s00234-006-0083-2.

Hedge, C., Powell, G., Sumner, P., 2018. The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. Behav. Res. Methods 50, 1166–1186. https://doi.org/10.3758/s13428-017-0935-1.

Herting, M.M., Gautam, P., Chen, Z., Mezher, A., Vetter, N.C., 2018. Test-retest reliability of longitudinal task-based fMRI: implications for developmental studies. Dev. Cogn. Neurosci. 33, 17–26. https://doi.org/10.1016/j.dcn.2017.07.001.

Infantolino, Z.P., Luking, K.R., Sauder, C.L., Curtin, J.J., Hajcak, G., 2018. Robust is not necessarily reliable: from within-subjects fMRI contrasts to between-subjects comparisons. Neuroimage 173, 146–152. https://doi.org/10.1016/j.neuroimage.2018.02.024.

Kim, H., 2011. Neural activity that predicts subsequent memory and forgetting: a meta-analysis of 74 fMRI studies. Neuroimage 54, 2446–2461. https://doi.org/10.1016/j.neuroimage.2010.09.045.

Koolschijn, P.C.M.P., Schel, M.A., de Rooij, M., Rombouts, S.A.R.B., Crone, E.A., 2011. A three-year longitudinal functional magnetic resonance imaging study of performance monitoring and test-retest reliability from childhood to early adulthood. J. Neurosci. 31, 4204–4212. https://doi.org/10.1523/JNEUROSCI.6415-10.2011.

Maril, A., Davis, P.E., Koo, J.J., Reggev, N., Zuckerman, M., Ehrenfeld, L., Mulkern, R.V., Waber, D.P., Rivkin, M.J., 2010. Developmental fMRI study of episodic verbal memory encoding in children. Neurology 75, 2110–2116. https://doi.org/10.1212/WNL.0b013e318201526e.

Mumford, J.A., Turner, B.O., Ashby, F.G., Poldrack, R.A., 2012. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. Neuroimage 59, 2636–2643. https://doi.org/10.1016/j.neuroimage.2011.08.076.

Nolden, S., Brod, G., Meyer, A.K., Fandakova, Y., Shing, Y.L., 2021. Neural correlates of successful memory encoding in kindergarten and early elementary school children: longitudinal trends and effects of schooling. Cereb. Cortex 31, 3764–3779. https://doi.org/10.1093/cercor/bhab046.

Ofen, N., 2012. The development of neural correlates for memory formation. Neurosci. Biobehav. Rev. 36, 1708–1717. https://doi.org/10.1016/j.neubiorev.2012.02.016.

Ofen, N., Chai, X.J., Schuil, K.D.I., Whitfield-Gabrieli, S., Gabrieli, J.D.E., 2012. The development of brain systems associated with successful memory retrieval of scenes. J. Neurosci. 32, 10012–10020. https://doi.org/10.1523/JNEUROSCI.1082-11.2012.

Ofen, N., Kao, Y.-C., Sokol-Hessner, P., Kim, H., Whitfield-Gabrieli, S., Gabrieli, J.D.E., 2007. Development of the declarative memory system in the human brain. Nat. Neurosci. 10, 1198–1205. https://doi.org/10.1038/nn1950.

Paller, K.A., Wagner, A.D., 2002. Observing the transformation of experience into memory. Trends Cogn. Sci. 6, 93–102. https://doi.org/10.1016/S1364-6613(00)01845-3.

Paz-Alonso, P.M., Ghetti, S., Donohue, S.E., Goodman, G.S., Bunge, S.A., 2008. Neurodevelopmental correlates of true and false recognition. Cereb. Cortex 18, 2208–2216. https://doi.org/10.1093/cercor/bhm246.

Plichta, M.M., Schwarz, A.J., Grimm, O., Morgen, K., Mier, D., Haddad, L., Gerdes, A.B. M., Sauer, C., Tost, H., Esslinger, C., Colman, P., Wilson, F., Kirsch, P., Meyer-Lindenberg, A., 2012. Test-retest reliability of evoked BOLD signals from a cognitive-emotive fMRI test battery. Neuroimage 60, 1746–1758. https://doi.org/10.1016/j.neuroimage.2012.01.129.

Power, J.D., Barnes, K. a, Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., 2012. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage 59, 2142–2154. https://doi.org/10.1016/j.neuroimage.2011.10.018.

Putcha, D., O'Keefe, K., Laviolette, P., O'Brien, J., Greve, D., Rentz, D.M., Locascio, J., Atri, A., Sperling, R., 2011. Reliability of functional magnetic resonance imaging associative encoding memory paradigms in non-demented elderly adults. Hum. Brain Mapp. 32, 2027–2044. https://doi.org/10.1002/hbm.21166.

Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R.J.A., Kahn, R.S., Ramsey, N.F., 2007. Test-retest reliability of fMRI activation during prosaccades and antisaccades. Neuroimage 36, 532–542. https://doi.org/10.1016/j.neuroimage.2007.03.061.

Rissman, J., Gazzaley, A., D'Esposito, M., 2004. Measuring functional connectivity during distinct stages of a cognitive task. Neuroimage 23, 752–763. https://doi.org/10.1016/j.neuroimage.2004.06.035.

Scoville, W.B., Milner, B., 1957. Loss of recent memory after bilateral hippocampal lesions. J. Neurol. Neurosurg. Psychiatry 20, 11–21. https://doi.org/10.1136/jnnp.20.1.11.

Shimamura, A.P., 1995. Memory and the prefrontal cortex. In: Structure and Functions of the Human Prefrontal Cortex, Annals of the New York Academy of Sciences. New York Academy of Sciences, New York, NY, US, pp. 151–159.

Shing, Y.L., Brehmer, Y., Heekeren, H.R., Bäckman, L., Lindenberger, U., 2016. Neural activation patterns of successful episodic encoding: reorganization during childhood, maintenance in old age. Dev. Cogn. Neurosci. 20, 59–69. https://doi.org/10.1016/j.dcn.2016.06.003.

Shing, Y.L., Werkle-Bergner, M., Brehmer, Y., Müller, V., Li, S.C., Lindenberger, U., 2010. Episodic memory across the lifespan: the contributions of associative and strategic components. Neurosci. Biobehav. Rev. 34, 1080–1091. https://doi.org/10.1016/j.neubiorev.2009.11.002.

Spaniol, J., Davidson, P.S.R., Kim, A.S.N., Han, H., Moscovitch, M., Grady, C.L., 2009. Event-related fMRI studies of episodic encoding and retrieval: meta-analyses using activation likelihood estimation. Neuropsychologia. https://doi.org/10.1016/j.neuropsychologia.2009.02.028.

Tang, L., Pruitt, P.J., Yu, Q., Homayouni, R., Daugherty, A.M., Damoiseaux, J.S., Ofen, N., 2020. Differential functional connectivity in anterior and posterior hippocampus supporting the development of memory formation. Front. Hum. Neurosci. 14, 1–16. https://doi.org/10.3389/fnhum.2020.00204.

Tang, L., Shafer, A.T., Ofen, N., 2018. Prefrontal cortex contributions to the development of memory formation. Cereb. Cortex 28, 3295–3308. https://doi.org/10.1093/cercor/bhx200.

Towgood, K., Barker, G.J., Caceres, A., Crum, W.R., Elwes, R.D.C., Costafreda, S.G., Mehta, M.A., Morris, R.G., von Oertzen, T.J., Richardson, M.P., 2015. Bringing memory fMRI to the clinic: comparison of seven memory fMRI protocols in temporal lobe epilepsy. Hum. Brain Mapp. 36, 1595–1608. https://doi.org/10.1002/hbm.22726.

Turner, B.O., Paul, E.J., Miller, M.B., Barbey, A.K., 2018. Small sample sizes reduce the replicability of task-based fMRI studies. Commun. Biol. 1. https://doi.org/10.1038/s42003-018-0073-z.

Vetter, N.C., Steding, J., Jurk, S., Ripke, S., Mennigen, E., Smolka, M.N., 2017. Reliability in adolescent fMRI within two years - a comparison of three tasks. Sci. Rep. 7, 1–11. https://doi.org/10.1038/s41598-017-02334-7.