


Test-retest reliability of hippocampal subfield volumes in a developmental sample: Implications for longitudinal developmental studies

Roya Homayouni^{1,2} | Qijing Yu^{1,2} | Sruthi Ramesh¹ | Lingfei Tang¹ |
Ana M. Daugherty^{1,2,3} | Noa Ofen^{1,2,4} 

¹Institute of Gerontology, Wayne State University, Detroit, MI, USA

²Department of Psychology, Wayne State University, Detroit, MI, USA

³Department of Psychiatry and Behavioral Neurosciences, School of Medicine, Wayne State University, Detroit, MI, USA

⁴Merrill Palmer Skillman Institute, Wayne State University, Detroit, MI, USA

Correspondence

Noa Ofen, Institute of Gerontology, Wayne State University, 257 Knapp, 87 E. Ferry, Detroit, MI 48202, USA.
Email: noa.ofen@wayne.edu

Present address

Sruthi Ramesh, Grossman School of Medicine, New York University, New York, NY, USA

Funding information

National Institutes of Health: Grant/Award Number: MH107512; Wayne State University Institute of Gerontology; Wayne State University Department of Psychology

Abstract

The hippocampus (Hc) is composed of cytoarchitecturally distinct subfields: dentate gyrus (DG), cornu ammonis sectors 1-3 (CA1-3), and subiculum. Limited evidence suggests differential maturation rates across the Hc subfields. While longitudinal studies are essential in demonstrating differential development of Hc subfields, a prerequisite for interpreting meaningful longitudinal effects is establishing test-retest consistency of Hc subfield volumes measured *in vivo* over time. Here, we examined test-retest consistency of Hc subfield volumes measured from structural MR images in two independent developmental samples. Sample One ($n = 28$, ages 7-20 years, $M = 12.64$, $SD = 3.35$) and Sample Two ($n = 28$, ages 7-17 years, $M = 11.72$, $SD = 2.88$) underwent MRI twice with a 1-month and a 2-year delay, respectively. High-resolution PD-TSE- T_2 -weighted MR images ($0.4 \times 0.4 \times 2 \text{ mm}^3$) were collected and manually traced using a longitudinal manual demarcation protocol. In both samples, we found excellent consistency of Hc subfield volumes between the two visits, assessed by two-way mixed intraclass correlation (ICC (3) single measures ≥ 0.87), and no difference between children and adolescents. The results further indicated that discrepancies between repeated measures were not related to Hc subfield volumes, or visit number. In addition to high consistency, with the applied longitudinal protocol, we detected significant variability in Hc subfield volume changes over the 2-year delay, implying high sensitivity of the method in detecting individual differences. Establishing unbiased, high longitudinal consistency of Hc subfield volume measurements optimizes statistical power of a hypothesis test and reduces standard error of the estimate, together improving external validity of the measures in constructing theoretical models of memory development.

KEYWORDS

brain development, hippocampus, longitudinal design, manual demarcation, measurement consistency, statistical power

Significance

Test-retest reliability provides internal validity for methods measuring brain volumes and it is an indispensable aspect of longitudinal designs. Longitudinal consistency is particularly important in assessing brain development as the attributes of interest and their correlates are subject to change over time. High reliability is critical to optimizing study design for sufficient statistical power of a hypothesis test and precision of parameter estimates. Here, using a longitudinal manual demarcation protocol, we demonstrate high longitudinal consistency of hippocampal subfield volume measurements in two developmental samples.

1 | INTRODUCTION

The hippocampus (Hc) is a fundamental neural substrate of episodic memory and is composed of cytoarchitecturally and functionally distinct subfields: dentate gyrus (DG), cornu ammonis sectors 1-3 (CA1-3), and subiculum regions (Duvernoy, 1988; Insausti & Amaral, 2012; Jones & McHugh, 2011). Available evidence in non-human primates suggests that structure of Hc subfields mature at different rates, and protracted development is reported for DG and CA3 (for review, see Lavenex & Lavenex, 2013). Human neuroimaging investigations lend further support for different developmental profiles of Hc subfield volumes between middle childhood and young adulthood, and age-related differences are primarily evidenced in the DG and CA3 volumes (Daugherty et al., 2016, 2017; Keresztes et al., 2017). Intriguingly, differential development of Hc subfields has been linked to the development of distinct memory functions such that delayed maturation of DG and CA3 suggested as contributing to the promotion of memory distinctiveness and complexity across development (Daugherty et al., 2017; Keresztes et al., 2017, 2018). With a few exceptions (Tamnes et al., 2014, 2018), the majority of available evidence regarding Hc subfields' structural and functional maturation comes from cross-sectional studies (Canada et al., 2018; Daugherty et al., 2017; Keresztes et al., 2017; Schlichting et al., 2017; Krosgrud et al., 2014; Lee et al., 2014) in which developmental effects cannot be truly assessed (Lindenberger et al., 2011). Hence, longitudinal studies are necessary to evaluate changes in Hc subfield volumes and potential contribution of mediators in the development of these regions. Critical to the study presented here, a prerequisite for interpreting meaningful longitudinal effects is a demonstrated test-retest reliability of Hc subfield volume measurements, yet few protocols have established this quality in longitudinal studies.

Test-retest reliability, the consistency of measurement over time, is assumed in all statistical tests of change in brain and behavior. Any measure, may it be cognitive performance or an MRI estimate of a brain structure, will vary between individuals and potentially within individuals over time. The variability of measures can be considered to include meaningful variance, indicative of development and its modifiers, and error variance. Hypotheses testing of developmental effects are sensitive to both sources of variance, and therefore measurement reliability is a prerequisite for valid interpretation of observed variability over time. In application to longitudinal hypothesis testing, reliability includes the consistency of measurement

with respect to both between- and within-person variability. In this manner, longitudinal hypothesis testing requires a demonstration of measurement consistency to interpret valid changes over time. However, to our knowledge, test-retest reliability of Hc subfield volume measurements has not been reported for most protocols applied to developmental samples (e.g., Tamnes et al., 2014; with one exception, Tamnes et al., 2018). Without evidence of test-retest consistency, interpretation of differences observed over time as reflecting developmental change will be confounded with error and replication will be difficult. Therefore, test-retest consistency can serve as required information when selecting adequate protocol to measure Hc subfield volumes in developmental studies.

In vivo estimates of regional brain measures obtained from MR images are sensitive to several sources of variance that are unique to the MRI methods and may contribute to measurement error in longitudinal studies. For example, drift in the scanning environment and variable positions within the scanner may result in differences in brain measures. Experiments that manipulated positions of participants found that procedures such as repositioning of a person in the scanner within the same session (Arshad et al., 2017; Brown et al., 2020) and scanning the same individual on a different day (Morey et al., 2010) may introduce error in the measurement (for review, see Brandmaier et al., 2018). Due to the small size of Hc subfields together with the vulnerability of the medial temporal lobe to signal loss (Olman et al., 2009), these sources of measurement error may obscure structural and functional estimates of longitudinal change, or exaggerate between-person differences.

Although test-retest reliability is required in any longitudinal study, the issue is particularly salient in developmental samples as the potential sources of measurement error may correlate with age, and therefore threaten the validity of conclusions about possible developmental changes. For example, a young 5-year-old has a 6%–10% smaller head circumference as compared to an older counterpart (Gaillard et al., 2001). Increased head size across development may modify MR signals differently over time within each subject (Kotsoni et al., 2006). Additionally, artifacts produced by motion or sinuses may correlate with age and further contribute to systematic bias in brain measures obtained from developmental samples (Dosenbach et al., 2017; Gaillard et al., 2001; Madan, 2018; Reuter et al., 2015). Last, measurement error in volumetric estimations that correlates with the size of a region may also correlate with age if the structure grows or shrinks across development, or differs between

persons across age, thus mutually confounding its interpretation (Schoemaker et al., 2016). Therefore, an assessment of test–retest reliability, and its correlates, is necessary to evaluate these potential sources of bias and to help determine if a protocol produces valid estimates of developmental changes in regional volumes.

Additional specific considerations for measurement errors are required for volumetric investigations of Hc subfields due to the differences in the size and unique morphological and anatomical properties of subfields (Marizzoni et al., 2015; Whelan et al., 2016; Yushkevich, Pluta, et al., 2015). Hypothesis test of mean change in Hc subfield volumes typically targets detecting the differential longitudinal changes across the subfields. To interpret differential developmental trajectories of Hc subfields, measurement reliability is assumed to be comparable across Hc subfields. Evaluating test–retest reliability is an indispensable step in ensuring that measurements are consistent over time, and comparable between individuals and across regions of interest (Putnick & Bornstein, 2016). Despite the importance of demonstrating test–retest reliability of MRI estimates of Hc subfield volumes in developmental samples, such endeavor is lacking in most segmentation protocols applied to developmental samples.

Hc subfields are defined by unique cytoarchitecture many of which are not evident *in vivo* using MRI. A gold standard for estimation of Hc subfield volumes is implementing a valid manual segmentation protocol based on commonly acquired high-resolution, TSE-PD-T₂-weighted MR images (Bender et al., 2018; Olsen et al., 2019; Wisse et al., 2016, 2020). In applying these procedures, additional recommended steps need to be implemented to reduce human rater bias (Nugent et al., 2007; Yushkevich, Amaral, et al., 2015). These additional steps are meant to eliminate the influence of possible changes in rater judgments when segmenting brain images obtained from the same person, to blind the rater to the time point of each scan, and to randomly distribute any remaining rater bias across individuals and across regions of interest. These recommendations are consistent with an understanding that even a measurement with the excellent reliability will still have error, but the error is to be randomly distributed so that it will not systematically bias the estimates of change over time or individual differences in development. These practical procedures improve the quality of data collected and any (semi) automatic segmentation protocols developed from it.

In the present study, we assessed the test–retest reliability of Hc subfield volumes in two healthy developmental samples ages 7–20 years. In Sample One, the first scan followed a short delay of 1 month, with the purpose of implementing a longitudinal study design while controlling for possible developmental change. Implementing such a design enabled us to attribute the differences observed between the two visits to changes in the imaging environment or other potential sources of volumetric fluctuations (i.e., hydration). In Sample Two, the follow-up to the first scan was after a 2-year delay as is typically implemented in longitudinal developmental studies. Our purpose was to assess the test–retest reliability of Hc regional volumes in the presence of potential developmental

changes, and whether our method is sensitive in detecting individual differences in the longitudinal change in Hc subfield volumes. Hc subfields were manually demarcated on high-resolution images using a published protocol (Bender et al., 2018) and implementing additional procedures for longitudinal data. In both samples, we assessed the test–retest reliability of volumetric measures of Hc subfields across two visits, and between children and adolescents. In Sample One, we further evaluated the possibility of measurement bias related to volumes of the subfield and the visit number. As additional step toward applying the protocol, we assessed the sensitivity of the method in detecting between-person differences in volume change over a 2-year delay (Sample Two). Last, we present the practical implications of high measurement reliability in study design to optimize statistical power when testing mean change and reduce standard error of the parameter estimates.

2 | MATERIAL AND METHODS

2.1 | Participants

Two samples of healthy, typically developing children were recruited from the Metro Detroit area as part of an ongoing longitudinal study. All participants underwent MRI twice in two separate visits. Delay between the scans in Sample One ($n = 28$, female = 14, ages 7–20 years, $Med = 12.15$, $M = 12.64$, $SD = 3.35$) was 1 month ($M = 30.21$ days, $SD = 3.63$) and delay between the scans in Sample Two ($n = 28$, female = 14, ages 7–17 years, $Med = 11.16$, $M = 11.72$, $SD = 2.88$) was 2 years ($M = 2.15$ years, $SD = 0.16$). Samples One and Two were comparable in terms of age and sex distribution (Figure S1). In presenting data from both samples, we use “Visit 1” and “Visit 2” when referring to the scans acquired in those respective visits. The sample size was calculated to achieve 80%–90% power ($N = 26$ – 35) to detect an expected ICC = 0.85 to be statistically significantly ($\alpha = 0.05$) greater than a minimum reliability of 0.60 (Bonett, 2002; Walter et al., 1998). To evaluate the possible age-related differences in ICC measures, the median split by age was used in defining the children and adolescents age-groups in both Sample One [children ($n = 14$, female = 6, ages 7.76–12.06 years, $M = 9.93$, $SD = 1.53$, delay: $M = 31.32$ days, $SD = 4.33$) and adolescents ($n = 14$, female = 8, ages 12.24–20.20 years, $M = 15.36$, $SD = 2.25$, delay: $M = 29.1$ days, $SD = 2.77$)] and Sample Two [children ($n = 14$, female = 5, ages 6.98–10.85 years, $M = 9.42$, $SD = 1.10$, delay: $M = 2.18$ years, $SD = 0.15$) and adolescents ($n = 14$, female = 9, ages 11.46–17.88 years, $M = 14.03$, $SD = 2.13$, delay: $M = 2.11$ years, $SD = 0.17$)]. All participants were screened for any potential developmental or neurological disorders or head trauma through phone interview or written questionnaires. For MRI session compatibility and safety issues, participants were excluded if they had any metallic implants, braces, or permanent retainers. Consent was obtained from all participants according to the procedures of the Wayne State University Institutional Review Board. Parental consent was acquired for participants who were under the age of 18 who also gave written or oral assent.

2.2 | Image acquisition and post-acquisition processing

High-resolution structural images were acquired, as a part of a 1-hr protocol, using a 32-channel head coil on a 3T Siemens Verio (Siemens Medical AG, Erlangen, Germany) scanner full-body magnet at Wayne State University. The high-resolution, proton density-weighted, turbo spin echo (PD-TSE) sequence was acquired perpendicular to the long axis of the Hc with the following parameters: voxel size = $0.4 \times 0.4 \times 2.0 \text{ mm}^3$ (30 slices); echo time (TE) = 17 ms; repetition time (TR) = 7,150 ms; flip angle = 120° ; pixel bandwidth = 96 Hz/pixel; turbo factor 11; and field of view (FOV) = $280 \times 512 \text{ mm}^2$. Due to possible morphological differences between left and right Hc (i.e., orientation and curvature), the left Hc was consistently used as the criterion structure in prescribing slice acquisition planes. This practice was implemented to maximize alignment similarities for data acquired in Visits 1 and 2 in both samples. In post-acquisition processing and using Analyze (v11.0) software, the intensity of all images was adjusted to 1,500 Hz and then inverted. These steps served the purpose of standardizing the intensity of image set and approximating the appearance of T_1 -weighted images for the ease of manual demarcation. The same data acquisition procedures and same processing steps were implemented in the analyses of data from Samples One and Two.

2.3 | Manual demarcation of Hc subfields

Hc subfields were manually demarcated on the images acquired perpendicular to the long axis of the Hc according to a reliable protocol that uses a geometric heuristic (Bender et al., 2018). The subfields were demarcated on contiguous coronal slices of the extant Hc body extending from the slice posterior to the uncus apex to the last slice

on which the lamina quadrigemina are visualized. The entorhinal cortex (EC), the main gateway to input the Hc subfields, was also traced for a total of six slices extending five slices anterior to the beginning of the Hc subfields body range (see Figure 1). Due to poor anatomical boundaries and to increase the reliability of volume measurements, subfields DG and CA3 were combined against the combined CA1 and CA2 regions. Within the context of development, the combination of DG and CA3, which exhibit protracted development, would allow us to capture the potential age-related changes. Three expert raters (R.H, Q.Y, and S.R) achieved high inter-rater reliability assessed by intraclass correlation coefficient, ICC (2) > 0.85 (for volumes from either right or left hemisphere) and ICC (2) > 0.9 (for the sum volumes from both hemispheres) across all regions of interest (ROIs; Shrout & Fleiss, 1979). Determining high inter-rater reliability assures that the raters are well trained, detail-oriented, and sufficiently experienced. All preprocessing steps and manual demarcations were done on a 21-in. digitizing tablet (Wacom Cintiq) with stylus in Analyze (v11.0) software.

2.4 | Randomization according to the longitudinal protocol

To safeguard against raters' biases related to the visit number, we implemented a procedure to blind raters for the information and to randomize bias. Deidentified data were coded randomly as A or B for the visit number. Pairs of tracers were randomly assigned to each participant's image set, and individual tracers were further randomized to slice assignment. This ensured that any rater-related bias was randomly distributed throughout the longitudinal data and any systematic error was avoided. The two images for each participant were then juxtaposed on the tablet, with placing A always on the left side and B on the right side of the screen. The tracers were required

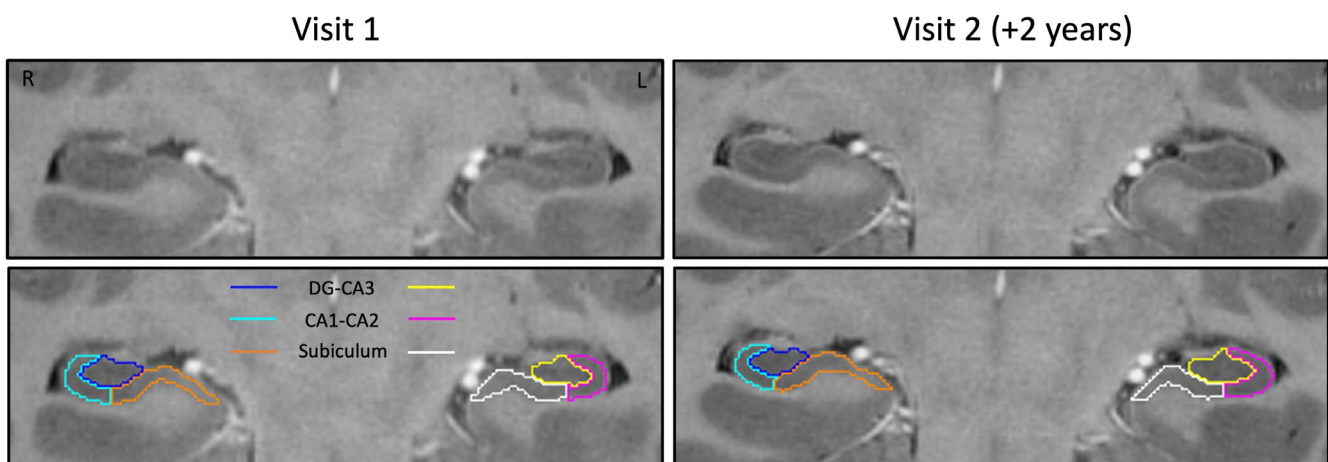


FIGURE 1 Manual demarcations of Hc subfields. Examples of high-resolution ($0.4 \times 0.4 \times 2 \text{ mm}^3$) PD-TSE- T_2 -weighted images from a 13-year-old male who was included in Sample Two and from which data were acquired in Visit 1 (left panels) and in a follow-up Visit 2 after 2 years delay (right panels). The intensity of the images is inverted for the ease of visualization. Bilateral Hc subfield demarcations are shown on the bottom panels, and structures without the demarcations are shown above for comparison. Color legends are depicted on the insert

to proceed the tracing slice by slice for both visits simultaneously and made a consistent decision for each visit in the event of any ambiguity (i.e., poor quality or artifact).

2.5 | Statistical analysis

2.5.1 | Test-retest reliability of ROIs using intraclass correlation coefficients (ICC3)

To assess the test-retest reliability of Hc subfield volume measurements, we utilized ICC. A two-way mixed-effect single ANOVA model that assumes nonindependent observation (ICC (3); Koo & Li, 2016; Shrout & Fleiss, 1979) was specifically conducted for Hc subfields. The repeated-measure design for internal consistency of the same individuals measured over time made by the same raters has an assumption of fixed measurement bias (Shrout & Fleiss, 1979); therefore, ICC (3,1) was computed for subfield volumetric measures (left, right, and total) comparing Visits 1 and 2 with a 95% confidence interval. An ICC value of 1 indicates perfect consistency in volume measurements between the repeated measures, and we expected a minimum acceptable ICC ≥ 0.85 , indicating approximately 15% measurement error tolerated in study design (Koo & Li, 2016). The available sample size was sufficient to provide at least 85% power to determine an ICC > 0.85 as statistically significantly different than a minimum ICC = 0.60.

2.5.2 | Bias evaluation using Bland-Altman plots

To determine if measurement error correlates with the volumetric measures of Hc subfield, we implemented Bland-Altman plots (Bland & Altman, 2007). This is of particular importance in the field of development, where Hc subfield volumes are subject to change, and volume-related bias may confound development-related changes. In addition to correlation analysis, we further assessed the bias toward the visit number using a one-sample *t*-test, where the volumetric discrepancies across the two visits (Visit 1– Visit 2) were compared against zero. False discovery rate (FDR, Benjamini & Hochberg, 1995) corrections were made for multiple comparisons for the 12 analyses (left and right hemispheres, and total volumes of both hemispheres for the four ROIs).

2.5.3 | Sensitivity analysis

There is the possibility that high reliability of volume measures is produced by the insensitivity of the method in capturing within-subject differences relative to between-subject variability. Due to potential developmental changes in Sample Two, we calculated the change scores of volume measures across two visits (Visit 1 – Visit 2). The sensitivity of the protocol to detect variability of change scores over 2-year delay was assessed by examining 95% confidence intervals

of variance estimated with bias-corrected accelerated bootstrapping (BCa 95% CI; 5,000 draws).

2.5.4 | Application to study planning: Simulated statistical power and standard error of measurement

Measurement reliability, as an index of the proportion of error, relates to sensitivity of a hypothesis test and the confidence intervals surrounding the parameter estimate; high measurement reliability provides high statistical power and low standard error. In investigating developmental changes in Hc subfield volumes, these are critical issues to be considered given the often small effect sizes of mean change, the large variability observed in developmental samples and the high inter-region correlations. Assuming that studies are designed intentionally to be representative of the population and to exclude sources of error, we estimated observed effect sizes in Sample Two over a 2-year delay as hypothetical magnitudes of mean change and as a function of measurement reliability (Williams & Zimmerman, 1989). The range of observed effect size estimates across ROIs was used in a power simulation (Faul et al., 2007, 2009) to determine the required sample size to achieve at least 85% power to detect the estimate to significance ($p < 0.05$) for measurements with different levels of reliability (ICC = 0.5, 0.6, 0.7, 0.8, 0.9, 0.99).

In addition to statistical significance, 95% confidence intervals are typically interpreted when considering the magnitude of the effect generalized to the population. The standard error of measurement (*SEM*) is calculated based on the observed standard deviation of estimated change and the measurement reliability [$SEM = \sigma\sqrt{(1 - ICC)}$], from which 95% confidence interval bounds are estimated [$\pm(1.96 \times SEM)$]. Small *SEM*, and therefore narrow 95% confidence intervals, indicates good precision of the estimated population parameter. We conducted simulations across the six levels of reliability (ICC = 0.5, 0.6, 0.7, 0.8, 0.9, 0.99) to calculate *SEM* as a function of standard deviation of change based on the variance observed in Sample Two.

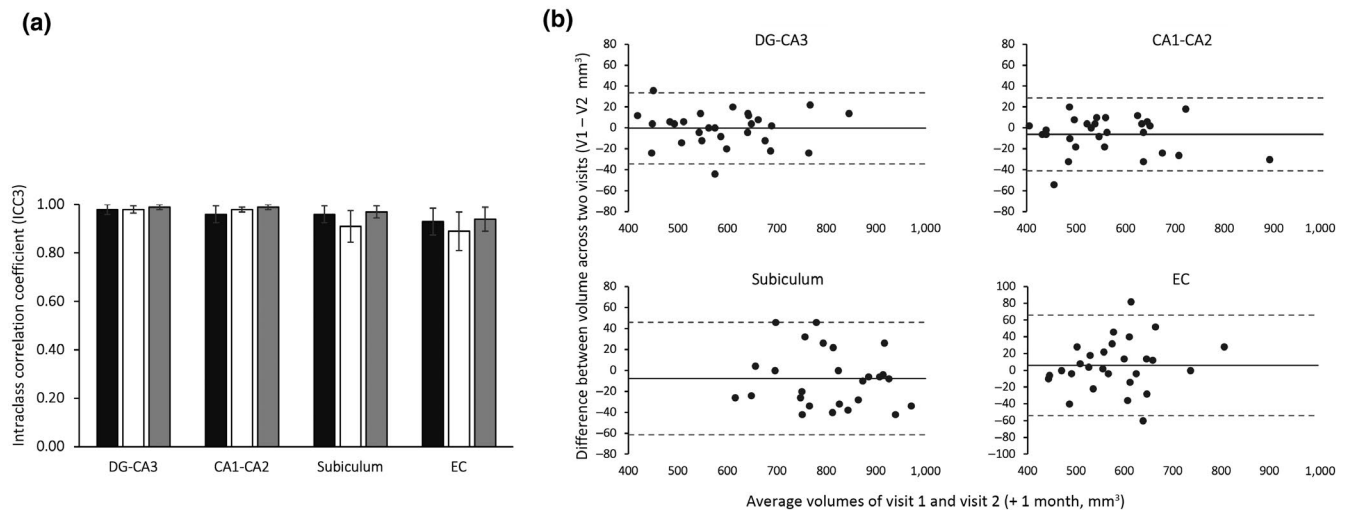
3 | RESULTS

3.1 | Sample One: Excellent consistency between visits with 1-month delay

All estimates of Hc subfield volumes had high test-retest reliability over a 1-month delay ranging between 0.89 and 0.99, as can be seen in the test-retest reliability coefficients presented in Table 1 and depicted in Figure 2a. Evaluation of measures of 95% CI indicated the hemispheres did not differ in the degree of measurement consistency. Observed volumetric discrepancies between two visits over 1-month delay did not significantly differ from zero in any ROI ($|t| < 1.8$, $p > 0.07$), suggesting no systematic bias toward the visit number. As the contribution of possible

TABLE 1 Intraclass correlation coefficients of (left, right, and total) Hc subfield and EC volumes for the whole sample and two age-groups over a 1-month delay

ROIs	ICC3 [CI] Sample One		
	Left	Right	Total
<i>Children and adolescents</i>			
DG-CA3	0.98 [0.95–0.99]	0.98 [0.95–0.99]	0.99 [0.97–0.99]
CA1-CA2	0.96 [0.92–0.98]	0.98 [0.97–0.99]	0.99 [0.97–0.99]
Subiculum	0.96 [0.91–0.98]	0.91 [0.82–0.96]	0.97 [0.94–0.99]
EC	0.93 [0.85–0.97]	0.89 [0.78–0.95]	0.94 [0.87–0.97]
<i>Children</i>			
DG-CA3	0.97 [0.98–0.99]	0.99 [0.96–0.99]	0.99 [0.97–0.99]
CA1-CA2	0.96 [0.90–0.99]	0.98 [0.95–0.99]	0.98 [0.96–0.99]
Subiculum	0.97 [0.89–0.98]	0.93 [0.80–0.97]	0.97 [0.92–0.99]
EC	0.91 [0.75–0.97]	0.84 [0.59–0.94]	0.91 [0.75–0.97]
<i>Adolescents</i>			
DG-CA3	0.97 [0.92–0.99]	0.94 [0.82–0.98]	0.97 [0.91–0.99]
CA1-CA2	0.93 [0.80–0.97]	0.98 [0.95–0.99]	0.98 [0.95–0.99]
Subiculum	0.95 [0.87–0.98]	0.90 [0.73–0.96]	0.96 [0.88–0.98]
EC	0.95 [0.85–0.98]	0.94 [0.83–0.98]	0.97 [0.91–0.99]

Sample One (one month delay)**FIGURE 2** (a) One-month delay consistency of Hc subfields and entorhinal cortex (EC) for left hemisphere (black), right hemisphere (white) and total volume (gray) are illustrated. The consistency of volumetric measures was equivalently high for all the ROIs and across hemispheres. Error bars represent the 95% confidence intervals. (b) The relationship between measurement error and average volume of Hc subfields and EC are provided. Within each plot, x-axis represents the average volumes across the two visits and y-axis represents the volumetric differences across two visits (Visits 1 and 2). Solid horizontal black line indicates the volume mean differences across two visits; horizontal dashed lines are 2 standard deviations above and below the mean. Difference scores are uniformly distributed across ROI sizes and within the 95% CI bands suggesting an unbiased volumetric estimation in respect to ROI sizes

developmental effects is expected to be minimal over a short, 1-month period, the stability of volume estimates combined with high test-retest consistency indicates that other factors such as changes in MRI environment between sessions or potential daily volumetric fluctuations have minimal impact on volume estimates of regions of interest.

Based on 95% CI, ICC3 measures were comparable between children and adolescents across all the ROIs (Table 1, Figure S2), and therefore the protocol has a negligible age-related bias. Although ICC3 coefficient was comparable between age-groups, we note that consistency of right EC volumes among children, although high, fell slightly below our target threshold: ICC3 = 0.84.

Bland–Altman plots were generated to evaluate the possibility of correlation between measurement consistency and ROI size (Figure 2b). Reviewing the Bland–Altman plots, we found that the majority of data fell within the 95% CI bands, uniformly distributed across ROI sizes and the volumetric difference between two visits did not correlate with ROI sizes (see Table 2). Notably, left EC measurement error correlation with size ($r = 0.42$, $p = 0.03$, $q = 0.004$) did not survive FDR correction.

3.2 | Sample Two: Excellent consistency between visits with 2-year delay

The results of test–retest reliability for Sample Two are depicted in Figure 3a. All the ROIs had high test–retest reliability (ICC3 = 0.87–0.96). Evaluation of 95% CI indicated that the consistency of volume measures was similar between hemispheres. Possible age-related bias was evaluated in comparing between children and adolescents. Based on 95% CI, ICC3 measures were comparable between two age-groups across all the ROIs (Table 3). Therefore, we replicated a

TABLE 2 Correlation coefficients for associations between average volumes and volumetric differences of Hc subfield and EC across two visits over a 1-month delay

	<i>r</i> [<i>p</i> value] Sample One		
	Left	Right	Total
ROIs	<i>Children and adolescents</i>		
DG-CA3	-0.12 [0.55]	0.06 [0.77]	-0.02 [0.90]
CA1-CA2	0.02 [0.92]	-0.12 [0.56]	-0.10 [0.63]
Subiculum	0.26 [0.18]	-0.16 [0.42]	-0.06 [0.77]
EC	0.42 [0.03]	-0.03 [0.88]	0.18 [0.35]

similar test–retest consistency between a 1-month and 2-year delay in two independent developmental samples.

3.3 | Sensitivity to detect the individual differences in longitudinal changes

Once high test–retest consistency was established, variability in the measure can be considered valid and would be used to test individual differences in development. Therefore, we aimed to assess if the reliable measures were sensitive to detect variability in change scores of volume measures across two visits over 2-year delay. The BCa 95% CI for the variance across ROIs in Sample Two showed that the variance of difference scores is different than zero in long-term delay (Figure 3b, Table 4) and comparable across regions. Therefore, the protocol was sensitive to individual differences in the magnitude (and direction) of change over the variable delay.

3.4 | Practical application of reliability assessments to longitudinal study planning

Based upon the estimated effect size of change and variability in change, we demonstrate how the information on measurement reliability can be used to plan a longitudinal study of development with at least 85% power to detect mean change to significance ($\alpha = 0.05$). As shown in Figure 4a, the statistical power of the test to detect mean change as significant is inversely proportional to measurement reliability. The practical implication of measurement reliability for determining the necessary sample size is most evident for tests of small effect size. The magnitude of change in regional Hc subfield and EC volumes over 2 years in Sample Two ranged $d = 0.2$ – 0.5 ; even

Sample Two (two years delay)

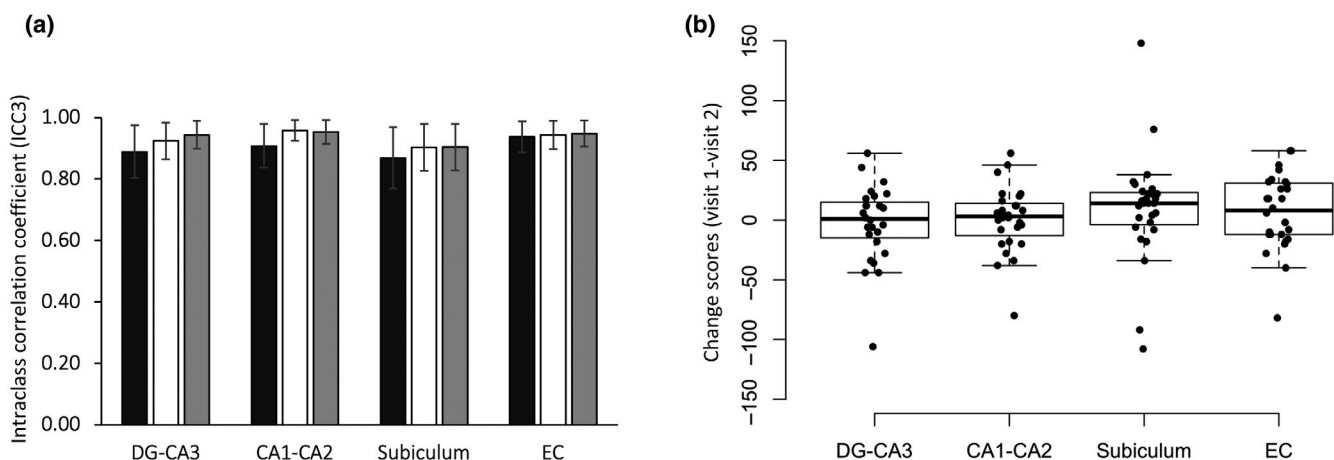


FIGURE 3 (a) Two-year delay consistency of Hc subfields and entorhinal cortex (EC) volumes are illustrated for left hemisphere (black), right hemisphere (white), and total volume (gray). The consistency of volumetric measures was equivalently high for all ROIs and across hemispheres. Error bars represent the 95% confidence intervals. (b) Sensitivity of the method in detecting individual differences in the change scores (Visits 1 and 2) is shown for the volumes of the ROIs over 2-year delay

TABLE 3 Intraclass correlation coefficients of (left, right, and total) Hc subfield and EC volumes for the whole sample and two age-groups over a 2-year delay

	ICC3 [CI] Sample Two		
	Left	Right	Total
ROIs	<i>Children and adolescents</i>		
DG-CA3	0.88 [0.77–0.95]	0.92 [0.84–0.96]	0.94 [0.88–0.97]
CA1-CA2	0.90 [0.81–0.96]	0.96 [0.91–0.98]	0.95 [0.90–0.98]
Subiculum	0.87 [0.73–0.94]	0.90 [0.80–0.95]	0.90 [0.80–0.96]
EC	0.94 [0.87–0.97]	0.94 [0.88–0.97]	0.95 [0.89–0.98]
	<i>Children</i>		
DG-CA3	0.84 [0.57–0.95]	0.87 [0.64–0.96]	0.91 [0.76–0.97]
CA1-CA2	0.88 [0.67–0.96]	0.92 [0.76–0.97]	0.93 [0.78–0.98]
Subiculum	0.88 [0.66–0.96]	0.88 [0.66–0.96]	0.91 [0.74–0.97]
EC	0.92 [0.77–0.97]	0.94 [0.83–0.98]	0.94 [0.81–0.98]
	<i>Adolescents</i>		
DG-CA3	0.96 [0.87–0.99]	0.95 [0.86–0.99]	0.97 [0.90–0.99]
CA1-CA2	0.95 [0.86–0.99]	0.98 [0.92–0.99]	0.97 [0.91–0.99]
Subiculum	0.86 [0.62–0.95]	0.91 [0.75–0.97]	0.90 [0.72–0.97]
EC	0.97 [0.91–0.99]	0.94 [0.83–0.98]	0.97 [0.90–0.99]

TABLE 4 Bootstrapped variance of change scores across (left, right, and total) Hc subfield and EC volumes over a 2-year delay

	Bootstrapped variance [CI], Sample Two		
	Left	Right	Total
ROIs	<i>Children and adolescents</i>		
DG-CA3	481.14 [229.25–726.23]	418.26 [203.62–623.81]	1,010.81 [454.26–1,629.94]
CA1-CA2	338.14 [187.97–470.76]	221.31 [107.70–327.12]	740.57 [360.11–1,102.66]
Subiculum	790.22 [434.20–1,112.16]	579.06 [275.23–884.59]	2,027.07 [818.93–3,257.13]
EC	400.66 [218.29–581.42]	347.66 [248.45–421.59]	983.24 [558.91–1,414.53]

with moderate effect sizes, the practical implication for sample recruitment demonstrates a benefit of high measurement reliability.

Similarly, high measurement reliability improves the precision of the parameter estimate (Figure 4b), illustrated by small standard error of measurement that would correspond to narrower 95% confidence intervals. Individuals are expected to differ in the magnitude (or direction) of change due to true variability in the population, and a highly reliable measure proportionally represents more true variability than error (Williams & Zimmerman, 1989). A study can effectively achieve comparable statistical power and confidence intervals of the estimate when using a low reliable measure by increasing sample size and limiting sample variability.

4 | DISCUSSION

To our knowledge, this is the first study to establish test-retest reliability of Hc subfield volumes using manual demarcation in a developmental sample. Simulating longitudinal study design, but controlling

for the possible effect of development, we first demonstrated excellent consistency of volumetric measures of Hc subfields and EC over 1-month delay. This indicated that change in factors involved in the scanning environment, such as head repositioning or alignments, does not substantially bias volume measures of Hc subfields and EC obtained with our protocol. Our findings also suggested that measurement error was independent of participant age, ROIs size, and visit number. Further, over a 2-year delay, we replicated excellent measurement consistency, and demonstrated high sensitivity of the method in detecting individual differences in change scores.

These results are consistent with previous findings assessing test-retest consistency of automated Hc subfields segmentation pipeline implemented in FreeSurfer in younger individuals (Tamnes et al., 2018) and healthy adults and neurodegenerative population (Brown et al., 2020; Mueller et al., 2018; Whelan et al., 2016; Worker et al., 2018). These studies report that volumetric estimations of Hc subfields obtained from FreeSurfer are highly consistent across different visits. Yet, the age-related bias of the Hc subfield volumes estimated by automated methods remains unclear. Particularly,

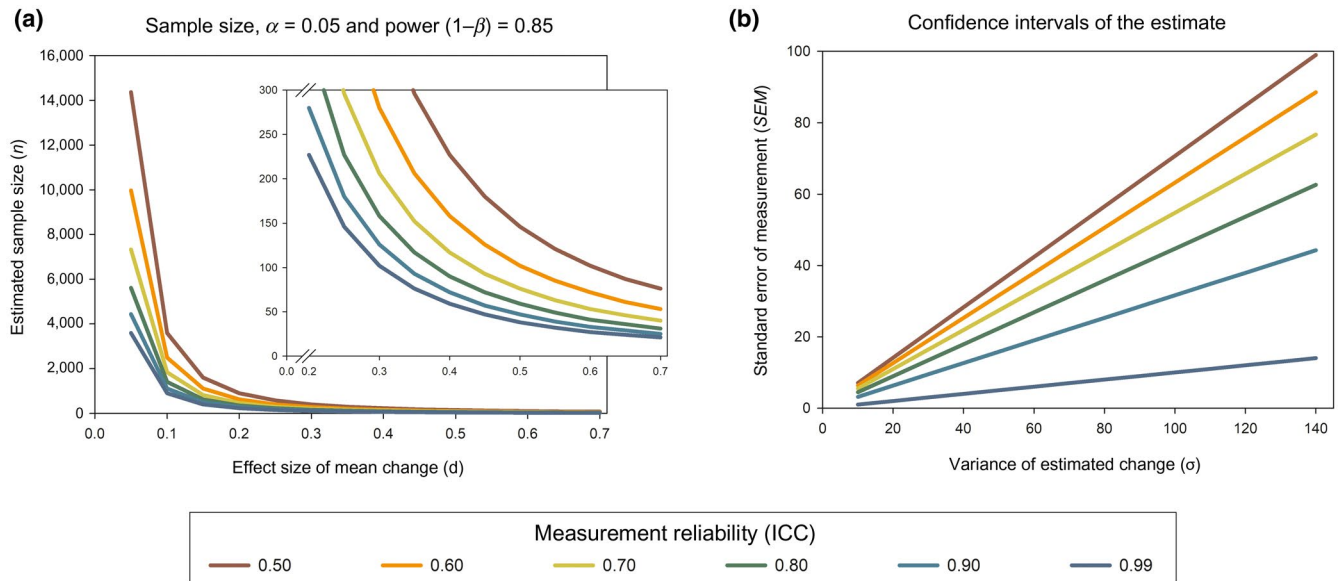


FIGURE 4 The practical applications of measurement reliability in optimizing the statistical power and confidence intervals of estimated mean change are illustrated. (a) The sample size required to achieve the minimum of 85% power ($\alpha = 0.05$) to detect an estimate of mean change to statistical significance ($p < 0.05$) is shown as a function of the true effect size and measurement reliability. The inset illustrates the range of effect sizes observed across regions of interest, approximately 0.2–0.5. High measurement reliability provides good sensitivity for estimated mean change in feasible sample sizes. (b) The standard error of measurement (SEM) in mean change, which is used to calculate 95% confidence intervals, is depicted as a function of observed variability in change and measurement reliability. Low SEM (narrower confidence interval bounds) indicates good precision of the parameter estimate, and high measurement reliability reduces SEM when variability in change is expected. The observed variability of regions of interest ranged from approximately 18 to 45

the automated segmentation tool is shown to be biased toward a different age-group or a different type of sample (for review, see Wenger et al., 2014), which is a critical consideration in developmental investigations. Besides, the Hc subfield segmentation with the aforementioned protocols has not been validated (Wisse, Biessels, & Geerlings, 2014), and therefore good test–retest reliability alone is insufficient to evaluate the protocol. Here, we applied a manual segmentation protocol that is reliable and has been validated in reference to known anatomy and in convergent studies across the life span (Bender et al., 2018), and provided additional critical information on the longitudinal test–retest consistency in a developmental sample.

A persistent challenge in the study of Hc subfield development and function, however, is the plethora of protocols that are highly variable and cannot be directly compared. There have been many empirical reviews that discuss these challenges (de Flores et al., 2019; DeKraker et al., 2020; Olsen et al., 2019; Yushkevich, Amaral, et al., 2015). Yet, even when the field at large comes to a consensus, accepted protocols will still need to consider test–retest consistency for their application to longitudinal study.

The reliability of a measure is not only the quality of its labels, but also the protocol by which the labels are applied. Manual demarcation, while considered a gold standard in *in vivo brain* segmentation, is also criticized for its vulnerability to the bias between and within raters (Tamnes et al., 2018); the strength and weakness both stem from subjective human judgment. The human rater is better equipped to apply a standardized protocol

to variable anatomy, or visualization, but this approach is also vulnerable to systematic bias. Here, we detail our intentional procedures to foster high reliability and maintain blind rating: in short, maximizing the strengths and mitigating the limitations of human judgment. Foremost, we assign two raters per case, each tracing a random subset of slices, but referring to each visit simultaneously. This promotes consistent rater decisions across slices and visits but reduces opportunities for systematic rater bias. Although between-rater agreement in our protocol was high, it is not without error, and the random assignment of raters randomly distributes error in the measurement, further mitigating statistical bias. These practical procedures, in addition to high inter-rater reliability of the manual segmentation protocol, resulted in an unbiased, high test–retest consistency over short and long delays and plausibly mitigated bias related to differences in MRI environment. As reliability is required for valid estimates of mean change and individual differences therein, procedures that promote good longitudinal measurement consistency are of equal importance to inter-rater agreement.

Important to the study of longitudinal change in Hc subfields, volumetric estimations acquired by high-resolution MRI are sensitive to individual differences, as shown with the current protocol. As reviewed above, Hc subfields exhibit differential developmental trajectories and their differential roles on memory development are yet to be explored (Daugherty et al., 2016, 2017; Jones & McHugh, 2011; Keresztes et al., 2017; Lavenex & Lavenex, 2013). Of relevance to characterizing cognitive development, a high-sensitive

method can capture nuance individual differences in the regions of interest which potentially contribute to the observed age-related differences in development of memory performance (Daugherty et al., 2017; Keresztes et al., 2018). Sensitive method for assessing Hc subfield volumes may help shed light on the functional specialization of the Hc subfields and facilitate the identification of the potential modifiers of individual developmental trajectories (i.e., genetic predisposition or socioeconomic status). Following the necessary assessment of test–retest reliability, the protocol can be applied in a larger sample to study mean change and individual differences over long periods.

We further demonstrated the practical implications of measurement reliability for study design to ensure statistical power of the hypothesis testing and narrow confidence intervals the parameter estimates. Statistical power is determined by total observed variance and can be modified by changing the true variance observed or the error variance (Zimmerman & Williams, 1986). A measure with low reliability will have large error variance and studies can compensate for this to achieve high statistical power by reducing the true variance by one of two approaches. The first option is to narrowly sample the population with strict selection criteria in order to minimize individual variability; however, this will reduce the external validity of the study and compounds concern about poor generalization to underrepresented groups. The second option is to increase the sample size and monopolize on regression toward this mean; as we demonstrate in the power and confidence interval simulations, the required sample sizes by this approach become intractable based on high cost of MRI and probable longitudinal attrition. An alternative approach is to improve statistical power by reducing error variance (Kanyongo et al., 2007), which has the added benefit of narrowing the 95% confidence intervals of an estimate to support interpretation of specific regional effects. Thus, investing in measurement protocols with high reliability is a cost-effective way to support longitudinal MRI studies with feasible sample sizes that are representative of the population. We suggest interpreting measurement reliability with respect to the expected true effect size, the observed variance, and the sample size rather than a rule-of-thumb threshold. The combination of the smallest true effect, largest observed variability, and lowest measurement reliability can be used to determine feasible study design. For example, a reliability coefficient of 0.7 indicates 30% of measurement variance is error—a value considered “good” by convention—yet the required sample size to achieve sufficient power and narrow confidence intervals may not be practical if the expected effect size is small and sample variability is large.

Overall, establishing test–retest reliability will assure that, in longitudinal studies, changes observed over time likely due to attributes of interest (e.g., age) rather than an artifact of measurement error. In developmental samples, accurate and reliable identification will facilitate research on the age-related changes of Hc subfield volumes, and will provide support for the internal validity of volumetric measures of Hc subfields in constructing theoretical models of memory development (Carr et al., 2010).

Reliability of the manual segmentation methods further extends to the quality of (semi) automated methods built from the protocol, which provides efficient data processing of large samples that are expected for longitudinal study. Finally, high measurement reliability by definition reduces sampled error, and by extension, improves statistical power of hypothesis tests, together this makes longitudinal studies with modest sample sizes more tractable (Zuo et al., 2019). This is of particular importance for the neuroimaging studies that are expensive in nature, and recruiting and retaining large samples is not always feasible.

4.1 | Limitations and future direction

While methodologically rigorous, the findings of this study should be interpreted in the context of several limitations. First, Hc subfield segmentation, due to the complexity of morphological and anatomical properties, was limited to the body of the Hc and adjacent EC, barring us to generalize to the Hc head and tail (Wisse et al., 2016). Second, our protocol does not provide the distinction across all the Hc subfields and combines smaller subfields into a single label: CA3–DG, CA1–CA2, and the subiculum regions. These decisions are made to improve the reliability of our manual segmentation protocol as the boundaries across the small subfields are poorly distinguishable at this image resolution and field strength (Iglesias et al., 2015; Marizzoni et al., 2015; Yushkevich, Amaral, et al., 2015). Third, we aimed to demonstrate the consistency and sensitivity of the measure in a developmental sample, and we cannot generalize the results to young childhood or across the adult life span. Finally, the current study was deliberately designed to examine possible sources of bias over variable delay and related to age; however, it falls short in detecting the specific source of the error measurement (i.e., effects of scan, day, or repositioning etc.). Because we found evidence of high test–retest consistency, other covariates of measurement error are likely to be small; but additional study of covariates would provide greater insight into sources of measurement error and options for statistical control in analysis (see Brandmaier et al., 2018).

5 | CONCLUSION

In this study, we established the longitudinal consistency of Hc subfield volumes in two independent developmental samples. We provided evidence on the unbiased reliability estimations of Hc subfield volumes which were independent of participant age, ROIs volumes, and visit number. We further showed that our methods were sensitive in capturing nuance individual differences of longitudinal changes. In conclusion, we present that methods applied are robust, reliable, and sensitive, which may increase the power of the hypothesis test and reduce standard error of measurement, particularly when the sample size or effect size are small, and high variability is expected. We suggest that current method, with further validation,

could support the use in measuring Hc subfield structural changes which potentially contribute the development of distinct aspects of episodic memory in children and adolescents.

DECLARATION OF TRANSPARENCY

The authors, reviewers and editors affirm that in accordance to the policies set by the *Journal of Neuroscience Research*, this manuscript presents an accurate and transparent account of the study being reported and that all critical details describing the methods and results are present.

ACKNOWLEDGMENTS

The authors thank Qin Yin, Bryn Thompson, Dana McCall, David Zhijian Chen, and Pavan Jella Kumar for the help with data collection.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHORS CONTRIBUTIONS

R.H.: *Conceptualization, Methodology, Investigation, Formal Analysis, Visualization, Writing Original Draft, Writing Review and Editing.* Q.J.: *Conceptualization, Methodology, Investigation, Formal Analysis, Visualization.* S.R.: *Investigation, Formal Analysis.* L.T.: *Conceptualization, Investigation, Formal Analysis.* A.M.D.: *Formal Analysis, Visualization, Writing Original Draft, Writing Review and Editing.* N.O.: *Conceptualization, Methodology, Investigation, Formal Analysis, Visualization, Writing Original Draft, Writing Review and Editing, Supervision, Funding Acquisition.*

PEER REVIEW

The peer review history for this article is available at <https://publons.com/publon/10.1002/jnr.24831>.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Noa Ofen  <https://orcid.org/0000-0003-3927-8408>

REFERENCES

Arshad, M., Stanley, J. A., & Raz, N. (2017). Test-retest reliability and concurrent validity of in vivo myelin content indices: Myelin water fraction and calibrated T1 w/T2 w image ratio. *Human Brain Mapping*, 38, 1780–1790.

Bender, A. R., Keresztes, A., Bodammer, N. C., Shing, Y. L., Werkle-Bergner, M., Daugherty, A. M., Yu, Q., Kühn, S., Lindenberger, U., & Raz, N. (2018). Optimization and validation of automated hippocampal subfield segmentation across the lifespan. *Human Brain Mapping*, 39, 916–931. <https://doi.org/10.1002/hbm.23891>

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300.

Bland, J. M., & Altman, D. G. (2007). Agreement between methods of measurement with multiple observations per individual. *Journal of Biopharmaceutical Statistics*, 17, 571–582. <https://doi.org/10.1080/10543400701329422>

Bonett, D. G. (2002). Sample size requirements for estimating intraclass correlations with desired precision. *Statistics in Medicine*, 21(9), 1331–1335. <https://doi.org/10.1002/sim.1108>

Brandmaier, A. M., Wenger, E., Bodammer, N. C., Kühn, S., Raz, N., & Lindenberger, U. (2018). Assessing reliability in neuroimaging research through intra-class effect decomposition (ICED). *eLife*, 7, e35718.

Brown, E. M., Pierce, M. E., Clark, D. C., Fischl, B. R., Iglesias, J. E., Milberg, W. P., McGlinchey, R. E., & Salat, D. H. (2020). Test-retest reliability of FreeSurfer automated hippocampal subfield segmentation within and across scanners. *Neuroimage*, 210, 116563. <https://doi.org/10.1016/j.neuroimage.2020.116563>

Canada, K. L., Ngo, C. T., Newcombe, N. S., Geng, F., & Riggins, T. (2018). It's all in the details: Relations between young children's developing pattern separation abilities and hippocampal subfield volumes. *Cerebral Cortex*, 29(8), 3427–3433.

Carr, V. A., Rissman, J., & Wagner, A. D. (2010). Imaging the human medial temporal lobe with high-resolution fMRI. *Neuron*, 65, 298–308. <https://doi.org/10.1016/j.neuron.2009.12.022>

Daugherty, A. M., Bender, A. R., Raz, N., & Ofen, N. (2016). Age differences in hippocampal subfield volumes from childhood to late adulthood. *Hippocampus*, 26(2), 220–228.

Daugherty, A. M., Flinn, R. W., & Ofen, N. (2017). Age-related differences in CA3-dentate gyrus volume uniquely linked to improvement in associative memory from childhood to adulthood. *Neuroimage*, 153, 75–85.

de Flores, R., Berron, D., Ding, S.-L., Ittyerah, R., Pluta, J. B., Xie, L., Adler, D. H., Robinson, J. L., Schuck, T., Trojanowski, J. Q., & Grossman, M. (2019). Characterization of hippocampal subfields using ex vivo MRI and histology data: Lessons for in vivo segmentation. *Hippocampus*, 30(6), 545–564. <https://doi.org/10.1002/hipo.23172>

DeKraker, J., Lau, J. C., Ferko, K. M., Khan, A. R., & Köhler, S. (2020). Hippocampal subfields revealed through unfolding and unsupervised clustering of laminar and morphological features in 3D BigBrain. *Neuroimage*, 206, 116328. <https://doi.org/10.1016/j.neuroimage.2019.116328>

Dosenbach, N. U. F., Koller, J. M., Earl, E. A., Miranda-Dominguez, O., Klein, R. L., Van, A. N., Snyder, A. Z., Nagel, B. J., Nigg, J. T., Nguyen, A. L., Wesevich, V., Greene, D. J., & Fair, D. A. (2017). Real-time motion analytics during brain MRI improve data quality and reduce costs. *NeuroImage*, 161, 80–93. <https://doi.org/10.1016/j.neuroimage.2017.08.025>

Duvernoy, H. M. (1988). *The human hippocampus: An atlas of applied anatomy*. JF Bergmann.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/BF03193146>

Gaillard, W. D., Grandin, C. B., & Xu, B. (2001). Developmental aspects of pediatric fMRI: Considerations for image acquisition, analysis, and interpretation. *Neuroimage*, 13(2), 239–249.

Iglesias, J. E., Augustinack, J. C., Nguyen, K., Player, C. M., Player, A., Wright, M., Roy, N., Frosch, M. P., McKee, A. C., Wald, L. L., Fischl, B., & Van Leemput, K. (2015). A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: Application

- to adaptive segmentation of in vivo MRI. *NeuroImage*, 115, 117–137. <https://doi.org/10.1016/j.neuroimage.2015.04.042>
- Insausti, R., & Amaral, D. G. (2012). Hippocampal formation. In J. K. Mai & G. Paxinos (Eds.), *The human nervous system* (pp. 896–942). Elsevier Inc.
- Jones, M. W., & McHugh, T. J. (2011). Updating hippocampal representations: CA2 joins the circuit. *Trends in Neurosciences*, 34, 526–535.
- Kanyongo, G. Y., Brook, G. P., Kyei-Blankson, L., & Gocmen, G. (2007). Reliability and statistical power: How measurement fallibility affects power and required sample sizes for several parametric and nonparametric statistics. *Journal of Modern Applied Statistical Methods*, 6(1), 81–90. <https://doi.org/10.22237/jmasm/1177992480>
- Keresztes, A., Bender, A. R., Bodammer, N. C., Lindenberger, U., Shing, Y. L., & Bergner, M. W. (2017). Hippocampal maturity promotes memory distinctiveness in childhood and adolescence. *Proceedings of the National Academy of Sciences of the United States of America*, 114, 9212–9217. <https://doi.org/10.1073/pnas.1710654114>
- Keresztes, A., Ngo, C. T., Lindenberger, U., Werkle-Bergner, M., & Newcombe, N. S. (2018). Hippocampal maturation drives memory from generalization to specificity. *Trends in Cognitive Sciences*, 22(2018), 676–686. <https://doi.org/10.1016/j.tics.2018.05.004>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Kotsoni, E., Byrd, D., & Casey, B. J. (2006). Special considerations for functional magnetic resonance imaging of pediatric populations. *Journal of Magnetic Resonance Imaging*, 23(6), 877–886. <https://doi.org/10.1002/jmri.20578>
- Krogsrud, S. K., Tamnes, C. K., Fjell, A. M., Amlie, I., Grydeland, H., Sultutvedt, U., Due-Tønnessen, P., Bjørnerud, A., Sølness, A. E., Håberg, A. K., Skrane, J., & Walhovd, K. B. (2014). Development of hippocampal subfield volumes from 4 to 22 years. *Human Brain Mapping*, 35, 5646–5657. <https://doi.org/10.1002/hbm.22576>
- Lavenex, P., & Lavenex, P. B. (2013). Building hippocampal circuits to learn and remember: Insights into the development of human memory. *Behavioral Brain Research*, 254, 8–21.
- Lee, J. K., Ekstrom, A. D., & Ghetti, S. (2014). Volume of hippocampal subfields and episodic memory in childhood and adolescence. *NeuroImage*, 94, 162–172. <https://doi.org/10.1016/j.neuroimage.2014.03.019>
- Lindenberger, U., von Oertzen, T., Ghisletta, P., & Hertzog, C. (2011). Cross-sectional age variance extraction: What's change got to do with it?. *Psychology and Aging*, 26(1), 34–47. <https://doi.org/10.1037/a0020525>
- Madan, C. R. (2018). Shape-related characteristics of age-related differences in subcortical structures. *Aging & Mental Health*, 23(7), 800–810. <https://doi.org/10.1080/13607863.2017.1421613>
- Marizzoni, M., Antelmi, L., Bosch, B., Bartrés-Faz, D., Müller, B. W., Wiltfang, J., Fiedler, U., Roccatagliata, L., Picco, A., Nobili, F., Blin, O., Bombois, S., Lopes, R., Sein, J., Ranjeva, J.-P., Didic, M., Gros-Dagnac, H., Payoux, P., Zoccatelli, G., ... Jovicich, J. (2015). Longitudinal reproducibility of automatically segmented hippocampal subfields: A multisite European 3T study on healthy elderly. *Human Brain Mapping*, 36, 3516–3527. <https://doi.org/10.1002/hbm.22859>
- Morey, R. A., Selgrade, E. S., Wagner, H. R., Huettel, S. A., Wang, L., & McCarthy, G. (2010). Scan-rescan reliability of subcortical brain volumes derived from automated segmentation. *Human Brain Mapping*, 31(11), 1751–1762. <https://doi.org/10.1002/hbm.20973>
- Mueller, S. G., Yushkevich, P. A., Das, S., Wang, L., Leemput, K. V., Iglesias, J. E., Alpert, K., Mezher, A., Ng, P., Paz, K., & Weiner, M. W. (2018). Systematic comparison of different techniques to measure hippocampal subfield volumes in ADNI2. *NeuroImage: Clinical*, 17, 1006–1018.
- Nugent III, T., Herman, D., Ordonez, A., Greenstein, D., Hayashi, K., Lenane, M., Clasen, L., Jung, D., Toga, A., & Giedd, J. (2007). Dynamic mapping of hippocampal development in childhood onset schizophrenia. *Schizophrenia Research*, 90(1-3), 62–70. <https://doi.org/10.1016/j.schres.2006.10.014>
- Olman, C. A., Davachi, L., & Inati, S. (2009). Distortion and signal loss in medial temporal lobe. *PLoS ONE*, 4(12), e8160. <https://doi.org/10.1371/journal.pone.0008160>
- Olsen, R. K., Carr, V. A., Daugherty, A. M., La Joie, R., Amaral, R. S. C., Amunts, K., Augustinack, J. C., Bakker, A., Bender, A. R., Berron, D., Boccardi, M., Bocchetta, M., Burggren, A. C., Chakravarty, M. M., Chételat, G., Flores, R., DeKraaker, J., Ding, S.-L., Geerlings, M. I., ... Wisse, L. E. M. (2019). Progress update from the hippocampal subfields group. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 11, 439–449. <https://doi.org/10.1016/j.dadm.2019.04.001>
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41, 71–90. <https://doi.org/10.1016/j.dr.2016.06.004>
- Reuter, M., Tisdall, M. D., Qureshi, A., Buckner, R. L., Van der Kouwe, A. J., & Fischl, B. (2015). Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *NeuroImage*, 107, 107–115. <https://doi.org/10.1016/j.neuroimage.2014.12.006>
- Schlichting, M. L., Guarino, K. F., Schapiro, A. C., Turk-Browne, N. B., & Preston, A. R. (2017). Hippocampal structure predicts statistical learning and associative inference abilities during development. *Journal of Cognitive Neuroscience*, 29, 37–51. https://doi.org/10.1162/jocn_a_01028
- Schoemaker, D., Buss, C., Head, K., Sandman, C. A., Davis, E. P., Chakravarty, M. M., Gauthier, S., & Pruessner, J. C. (2016). Hippocampus and amygdala volumes from magnetic resonance images in children: Assessing accuracy of FreeSurfer and FSL against manual segmentation. *NeuroImage*, 129, 1–14. <https://doi.org/10.1016/j.neuroimage.2016.01.038>
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428.
- Tamnes, C. K., Bos, M. G. N., van de Kamp, F. C., Peters, S., & Crone, E. A. (2018). Longitudinal development of hippocampal subregions from childhood to adulthood. *Developmental Cognitive Neuroscience*, 30, 212–222. <https://doi.org/10.1016/j.dcn.2018.03.009>
- Tamnes, C. K., Walhovd, K. B., Engvig, A., Grydeland, H., Krogsrud, S. K., Østby, Y., Holland, D., Dale, A. M., & Fjell, A. M. (2014). Regional hippocampal volumes and development predict learning and memory. *Developmental Neuroscience*, 36, 161–174. <https://doi.org/10.1159/000362445>
- Walter, S. D., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine*, 17, 101–110. [https://doi.org/10.1002/\(SICI\)1097-0258\(19980115\)17:1<101:AID-SIM727>3.0.CO;2-E](https://doi.org/10.1002/(SICI)1097-0258(19980115)17:1<101:AID-SIM727>3.0.CO;2-E)
- Wenger, E., Mårtensson, J., Noack, H., Bodammer, N. C., Kühn, S., Schaefer, S., Heinze, H.-J., Düzel, E., Bäckman, L., Lindenberger, U., & Lövdén, M. (2014). Comparing manual and automatic segmentation of hippocampal volumes: Reliability and validity issues in younger and older brains. *Human Brain Mapping*, 35, 4236–4248. <https://doi.org/10.1002/hbm.22473>
- Whelan, C. D., Hibar, D. P., van Velzen, L. S., Zannas, A. S., Carrillo-Roa, T., McMahon, K., Prasad, G., Kelly, S., Faskowitz, J., deZubiracay, G., Iglesias, J. E., van Erp, T. G. M., Frodl, T., Martin, N. G., Wright, M. J., Jahanshad, N., Schmaal, L., Sämann, P. G., & Thompson, P. M. (2016). Heritability and reliability of automatically segmented human hippocampal formation subregions. *NeuroImage*, 128, 125–137. <https://doi.org/10.1016/j.neuroimage.2015.12.039>

- Williams, R. H., & Zimmerman, D. W. (1989). Statistical power analysis and reliability of measurement. *Journal of General Psychology*, *116*(4), 359–369. <https://doi.org/10.1080/00221309.1989.9921123>
- Wisse, L. E. M., Biessels, G. J., & Geerlings, M. I. (2014). A critical appraisal of the hippocampal subfield segmentation package in FreeSurfer. *Frontiers in Aging Neuroscience*, *6*, 261. <https://doi.org/10.3389/fnagi.2014.00261>
- Wisse, L. E. M., Chételat, G., Daugherty, A. M., de Flores, R., la Joie, R., Mueller, S. G., Stark, C. E., Wang, L., Yushkevich, P. A., Berron, D., Raz, N., Bakker, A., Olsen, R. K., & Carr, V. A. (2020). Hippocampal subfield volumetry from structural isotropic 1 mm³ MRI scans: A note of caution. *Human Brain Mapping*, *42*(2), 539–550.
- Wisse, L. E. M., Daugherty, A. M., Amaral, R. S. C., Berron, D., Carr, V. A., Ekstrom, A. D., Kanel, P., Kerchner, G. A., Mueller, S. G., Pluta, J., Stark, C. E., Steve, T., Wang, L., Yassa, M. A., Yushkevich, P. A., & La Joie, R. (2016). P2–060: A harmonized protocol for medial temporal lobe subfield segmentation: Initial results of the 3-tesla protocol for the hippocampal body. *Alzheimer's & Dementia*, *12*, P631. <https://doi.org/10.1016/j.jalz.2016.06.1265>
- Worker, A., Dima, D., Combes, A., Crum, W. R., Streffer, J., Einstein, S., Mehta, M. A., Barker, G. J., Williams, S. C., & O'daly, O. (2018). Test-retest reliability and longitudinal analysis of automated hippocampal subregion volumes in healthy ageing and Alzheimer's disease populations. *Human Brain Mapping*, *39*(4), 1743–1754.
- Yushkevich, P. A., Amaral, R. S. C., Augustinack, J. C., Bender, A. R., Bernstein, J. D., Boccardi, M., Burggren, A. C., Carr, V. A., Chakravarty, M. M., & Chételat, G. (2015). Quantitative comparison of 21 protocols for labeling hippocampal subfields and parahippocampal subregions in in vivo MRI: Towards a harmonized segmentation protocol. *Neuroimage*, *111*, 526–541.
- Yushkevich, P. A., Pluta, J. B., Wang, H., Xie, L., Ding, S. L., Gertje, E. C., & Wolk, D. A. (2015). Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. *Human Brain Mapping*, *36*, 258–287.
- Zimmerman, D. W., & Williams, R. H. (1986). Note on the reliability of experimental measures and the power of significance tests. *Psychological Bulletin*, *100*(1), 123–124.
- Zuo, X. N., Xu, T., & Milham, M. P. (2019). Harnessing reliability for neuroscience research. *Nature Human Behaviour*, *3*(8), 768–771.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

FIGURE S1 Samples characteristics plotted as participants distribution by age and sex (female = red, male = blue) for Sample One (one-month delay, left panel) and for Sample Two (two-years delay, right panel). Each sample consisted of 28 participants and none of the participants were included in more than one of the samples. Dots at two ends of each line represent participant age at Visit 1 and Visit 2 and the length of the lines indicates the delay between the two visits

FIGURE S2 Consistency of volumetric measures over a one-month delay did not differ by age. Consistency, measured by intraclass correlation coefficients (ICC3), is depicted separately for subsample of children (turquoise) and adolescents (purple) defined based on median split by age. The consistency of volumetric measures was equivalently high in all regions of interest (left, right, and total volumes of Hc subfields and EC) in both subsamples. Figure depicting analyses from Sample One, similar pattern was found for sample 2 (see text). Error bars represent the 95% confidence intervals. CA, cornu ammonis; DG, dentate gyrus; EC, entorhinal cortex

Transparent Science Questionnaire for Authors
Transparent Peer Review Report

How to cite this article: Homayouni R, Yu Q, Ramesh S, Tang L, Daugherty AM, Ofen N. Test-retest reliability of hippocampal subfield volumes in a developmental sample: Implications for longitudinal developmental studies. *J Neurosci Res*. 2021;99:2327–2339. <https://doi.org/10.1002/jnr.24831>