Hippocampal Structure Predicts Statistical Learning and Associative Inference Abilities during Development

Margaret L. Schlichting[1], Katharine F. Guarino[1], Anna C. Schapiro[2], Nicholas B. Turk-Browne[3], and Alison R. Preston[1]

Abstract

■ Despite the importance of learning and remembering across the lifespan, little is known about how the episodic memory system develops to support the extraction of associative structure from the environment. Here, we relate individual differences in volumes along the hippocampal long axis to performance on statistical learning and associative inference tasks—both of which require encoding associations that span multiple episodes—in a developmental sample ranging from ages 6 to 30 years. Relating age to volume, we found dissociable patterns across the hippocampal long axis, with opposite nonlinear volume changes in the head and body. These structural differences were paralleled by performance gains across the age range on both tasks, suggesting improvements in the cross-episode binding ability from childhood to adulthood. Controlling for age, we also found that smaller hippocampal heads were associated with superior behavioral performance on both tasks, consistent with this region's hypothesized role in forming generalized codes spanning events. Collectively, these results highlight the importance of examining hippocampal development as a function of position along the hippocampal axis and suggest that the hippocampal head is particularly important in encoding associative structure across development. ■

INTRODUCTION

Much is known about how the mature hippocampus (HPC) gives rise to our ability to learn and remember; however, whether the improvements in memory behavior observed across development (Brainerd, Holliday, & Reyna, 2004; Billingsley, Smith, & McAndrews, 2002; for a review, see Ghetti & Bunge, 2012) can be attributed to HPC structural changes has not been well characterized. Here, we investigate the relationship between developmental differences in HPC structure and behavioral performance on two tasks that require forming associations across time. Dynamically extracting such associative regularities is thought to be critical to cognitive function across development, supporting such fundamental abilities as language learning and knowledge acquisition (Frost, Siegelman, Narkiss, & Afek, 2013; Aslin & Newport, 2012).

Prior research in adults has related HPC task engagement to the formation and retrieval of memories for associations experienced not only within (Schacter & Wagner, 1999) but also across events (Schlichting, Zeithamova, & Preston, 2014; Schapiro, Kustner, & Turk-Browne, 2012; Zeithamova, Dominick, & Preston, 2012; Zeithamova & Preston, 2010). For example, a host of empirical work has documented greater HPC engagement when an episode is subsequently remembered relative to forgotten (Paller & Wagner, 2002), implicating HPC in memory formation. However, existing research also suggests that even stable neural characteristics account for variability in memory performance. For example, a number of studies have linked HPC volumes to memory for individual episodes (Demaster, Pathman, Lee, & Ghetti, 2013; Poppenk & Moscovitch, 2011; Maguire, Woollett, & Spiers, 2006).

In this study, we investigate the relationship between HPC volumes and performance on two memory tasks—statistical learning and associative inference—that require participants to encode cross-episode associative structure. In the statistical learning task (Schapiro, Gregory, Landau, McCloskey, & Turk-Browne, 2014; Fiser & Aslin, 2002), participants extract associative information from a continuous stream of shapes by detecting repeated sequences of specific shapes. In the associative inference task (Preston, Shrager, Dudukovic, & Gabrieli, 2004), participants encode associations that overlap via a shared item and are later tested on their ability to link the related pairs.

Although much is known about how HPC supports memory in the adult brain, little work has investigated its development beyond early childhood (Gómez & Edgin, 2015; Lavenex & Banta Lavenex, 2013). Research on age-related structural differences has produced conflicting findings, with some reporting overall HPC volume increases with age (Østby et al., 2009) and others reporting no change at all beyond early childhood (Yurgelun-Todd,

[1]The University of Texas at Austin, [2]Beth Israel Deaconess Medical Center/Harvard Medical School, [3]Princeton University

Journal of Cognitive Neuroscience 29:1, pp. 37–51
doi:10.1162/jocn_a_01028

Killgore, & Cintron, 2003; Giedd et al., 1996). However, such apparent inconsistencies might arise from differences in HPC subregional development (Daugherty, Yu, Flinn, & Ofen, 2015; Lee, Ekstrom, & Ghetti, 2014; Demaster et al., 2013; Gogtay et al., 2006), as volume has been shown to decrease in anterior and increase in more posterior HPC regions (Demaster et al., 2013; Gogtay et al., 2006).

These subtle structural changes may lead to important developmental shifts in behavior. Adult HPC function is known to be highly heterogeneous, both across subfields (cornu ammonis [CA] fields, dentate gyrus [DG], and subiculum) and along its anterior–posterior axis across head, body, and tail (Strange, Witter, Lein, & Moser, 2014; Poppenk & Moscovitch, 2011). For example, in adults, posterior HPC volume positively predicts memory, whereas anterior HPC volume negatively predicts memory (Demaster et al., 2013; Poppenk & Moscovitch, 2011; Maguire et al., 2000, 2006). This dissociation in the volume–performance relationship may be a result of the differential granularity of the memory representation coded by anterior (i.e., head) and posterior (body, tail) regions: Whereas posterior HPC is thought to encode specific details, anterior forms generalized codes (Collin, Milivojevic, & Doeller, 2015; Schlichting, Mumford, & Preston, 2015). Here, we test the central hypothesis that anterior HPC volume in participants aged 6–30 years would uniquely track cross-episodic binding behavior, which relies on the formation of generalized memory representations. We also predicted that volume–performance relationships would interact with age, reflecting increased ability to bind information across episodes in adults.

METHODS

Participants

Ninety volunteers participated in the experiment across child (6–11 years; $n = 31$), adolescent (12–17 years; $n = 25$), and adult (18–30 years; $n = 34$) age groups. The consent/assent process was carried out using age-appropriate language in accordance with an experimental protocol approved by the institutional review board at the University of Texas at Austin. Permission was obtained from the parent of participants under age 18. All participants received monetary compensation and a small prize for their involvement in the study.

Participants were screened for psychiatric conditions using the Child Behavior Checklist (CBCL; completed by the parent/guardian of participants aged 6–17; Achenbach, 1991) and the Symptom Checklist 90-Revised (SCL-90-R; adults; Derogatis, 1977). IQ was assessed using the Wechsler Abbreviated Scale of Intelligence, Second Edition (WASI-II; Wechsler, 2011). The intelligence measure of interest was the full-scale IQ composite score (FSIQ-2), which includes vocabulary and matrix reasoning subtests.

From the original group of 90 participants, individuals were excluded from all subsequent analyses if they met either of the following criteria: (1) SCL-90-R score above the normal range (greater than 1 *SD* above the mean of a normative sample; $n = 9$ adults) or CBCL score in the clinical range ($n = 1$ child; $n = 1$ adolescent); or (2) presence of a psychiatric condition ($n = 1$ adult). No participants scored below our inclusion threshold for IQ (>2 *SD* below the mean). This initial round of exclusions yielded a group of 30 children (16 girls; age: range = 6.00–11.83 years, mean ± *SEM* = 9.30 ± 0.32; FSIQ-2: range = 84–142, 119.00 ± 2.38), 24 adolescents (10 girls; age: 12.00–17.25 years, 14.15 ± 0.33 years; FSIQ-2: 92–142, 113.00 ± 2.63), and 24 adults (14 women, age: 18.58–29.50 years, 24.11 ± 0.71 years; FSIQ-2: 92–135, 113.54 ± 2.40) available for inclusion in our analyses. From this group, we identified participants with acceptable behavioral and/or neural data for each analysis using the exclusion criteria below.

Exclusions for Behavioral Analyses

Participants were excluded from the statistical learning analysis if they (1) completed an earlier version of the task ($n = 2$ children, $n = 2$ adolescents, $n = 2$ adults) or (2) failed to perform the task as directed (e.g., responding before viewing both response options; $n = 5$ children). In total, 67 participants were included for behavioral analyses of statistical learning ($n = 23$ children, $n = 22$ adolescents, $n = 22$ adults).

Participants were excluded from the associative inference analysis if they either failed to perform above chance on the final direct pair test ($n = 6$ children; $n = 2$ adolescents) or were unable to understand the task instructions ($n = 1$ child). These exclusions yielded a total n of 69 for behavioral analyses of associative inference ($n = 23$ children, $n = 22$ adolescents, $n = 24$ adults).

Exclusions for Age–Volume Relationships

Participants were excluded from the neural analyses if they met any of the following criteria: (1) did not complete the MRI portion ($n = 3$ children, $n = 4$ adults), (2) MRI data not of acceptable quality ($n = 3$ children; see MR Data Acquisition and Analyses), or (3) handedness concerns ($n = 1$ adolescent). The final sample for the volume analyses included a total of 68 right-handed participants (24 children, 23 adolescents, 21 adults).

Exclusions for Volume–Behavior Relationships

Participants were included in the volume–behavior analyses if they met the inclusion criteria described above for both the behavioral and age–volume analyses. In total, 60 participants were included for analyses related to the statistical learning task ($n = 20$ children, $n = 21$

adolescents, n = 19 adults), and 62 were included for the associative inference task (n = 20 children, n = 21 adolescents, n = 21 adults).

Experiment Overview

On the first of two visits, participants were exposed to the MRI environment using a mock scanner, completed paper-based screening measures (CBCL or SCL-90-R; WASI-II), and performed a battery of cognitive tasks, including associative inference and statistical learning tasks. The order of tasks was fixed across participants: associative inference, Iowa gambling (Bechara, Damasio, Damasio, & Anderson, 1994), statistical learning, relational reasoning (modified version of Raven's Progressive Matrices; Crone et al., 2009), and a relational integration task (as described in Wendelken & Bunge, 2010). Because this study focuses on associative memory formation, data from Iowa gambling, relational reasoning, and relational integration tasks are not reported. MRI scanning took place during the second visit.

Statistical Learning Task

Materials. Participants completed a shape version of the statistical learning task as described in a prior study (Figure 1A; Schapiro et al., 2014). Stimuli were 12 novel shapes (Fiser & Aslin, 2001) organized into four groups of three shapes or triplets. For each participant, shapes were assigned randomly and without replacement to a specific triplet and position within the triplet (first, second, third). For both the statistical learning and associative inference tasks, we used novel rather than familiar stimuli to avoid differences in preexisting knowledge across ages.

Procedure. During the initial familiarization phase, participants viewed shapes presented one at a time for 0.5 sec with an ISI of 0.5 sec (both for items within and for items spanning triplets, resulting in a continuous sequence). Each of the four triplets was viewed 24 times, for a total of 288 item presentations. The presentation order of specific triplets was randomized across participants, with the added constraints that neither (a) a single triplet (e.g., ABC) nor (b) a pair of triplets (ABCDEF) could be repeated in immediate succession. These constraints were imposed to minimize the likelihood of explicit detection of the sequences. Because the sequence was continuous, the only cue to the existence of triplets was the higher transition probability within versus between triplets. That is, the probability of transitioning from a given item to the next within triplet (e.g., A → B) is 1, whereas the transition probability for a pair of items that span triplets (e.g., C → D) is approximately 1/3. Thus, participants must extract these statistical regularities to learn the structure of the sequence. Items within a triplet were always presented sequentially in the same order within the larger familiarization sequence. No reference was made to participants about the triplets or embedded sequences.

After the familiarization phase, participants completed a self-paced two-alternative forced-choice test for their knowledge of triplet sequences (i.e., a triplet discrimination task). During the test, participants were asked to judge the relative familiarity of two sequences of three shapes. One of the sequences consisted of a triplet viewed during the familiarization phase whereas the other choice was one of four foils. Text instructing participants to indicate the more familiar sequence with a key press was displayed on the screen; the researcher also instructed participants verbally using age-appropriate language to ensure understanding (e.g., for young children: "Watch the three pictures as they show up on each side the screen. Which group feels more familiar or like you have seen it before?"). Critically, because all items were presented with equal frequency during familiarization

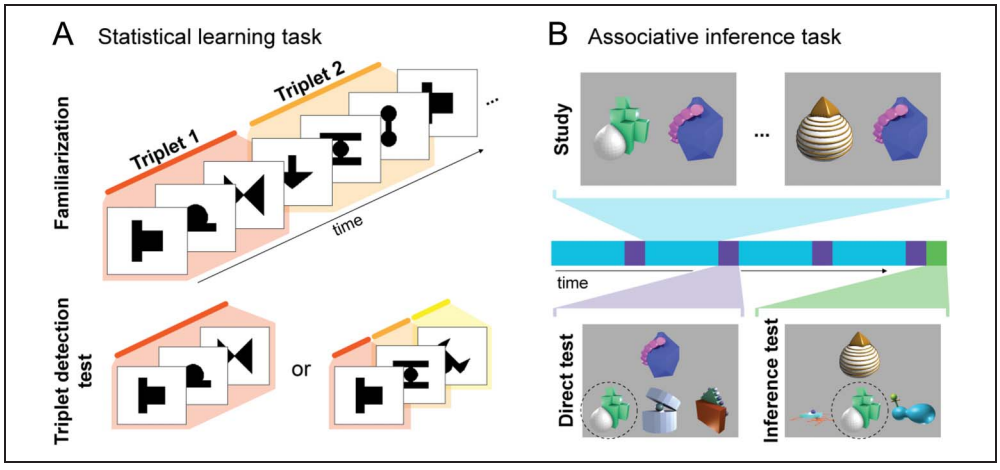

Figure 1. Behavioral task designs. (A) Statistical learning task. During familiarization, participants viewed a sequence of shapes organized into triplets (indicated with shading; not shown to participants). Participants then completed a triplet discrimination test in which they selected which of two sequences was more familiar. (B) Associative inference task. Participants learned overlapping pairs across four study–test iterations (blue, purple). During the tests, participants selected which of the bottom three choice objects was studied with the top cue object (correct answer circled for display). Following learning, participants completed an inference test (green) in which they linked indirectly related items through their common association with an overlapping object.

and test, any familiarity preference for the triplet indicates statistical learning of the transition probabilities. The foils consisted of shapes from three different triplets, which never appeared consecutively during the familiarization phase. For the foils, shapes retained their assigned positions (first, second, third), but triplet membership was shuffled. Each of the four triplets was paired with each of the four foils once in each half of the test, such that each triplet was tested eight times. For each test trial, the first option (target or foil, counterbalanced) appeared on one side of the screen (left or right, randomized) with the same timing as the familiarization phase. Following a 1-sec delay, the second option appeared on the screen. Participants had the option to view each trial again before making their response.

Analysis of behavioral data. Accuracy on the test was computed for each participant; a one-way ANOVA tested for differences in triplet discrimination performance across Age groups.

Associative Inference Task

Materials. Participants completed a modified version of the associative inference task (Preston et al., 2004). Stimuli were 45 novel objects (Schlichting et al., 2015) created using Blender (www.blender.org), a subset of which were adapted from a prior study (Hsu, Schlichting, & Thompson-Schill, 2014). Novel objects were arranged into 15 groups of three, termed ABC triads. ABC triads were presented as overlapping AB and BC pairs, with the B item shared between pairs (Figure 1B, top). That is, AB pairs consisted of two novel objects, A and B; the B object was then later paired with a new novel object C to form a BC pair. Because of the relatively small number of learning opportunities for each pair (four presentations per pair; see below), we reasoned that (a) differences in the visual similarity of paired items (e.g., having a similar color) and/or (b) differences in study order across participants might impact overall task difficulty. For this reason and because the primary goal of this study was to index how individual differences in memory relate to structural measures (Carlson & Moses, 2013), we equated difficulty to the extent possible by fixing triad assignment and presentation order across participants.

Procedure. Participants first learned the AB and BC pairs in four study–test alternations. During each study phase (Figure 1B, blue), novel object pairs were presented on the screen for 3.5 sec with an ISI of 0.5 sec. Each of the 30 AB and BC pairs was presented once. For AB pairs, A objects always appeared on the left side of the screen; for BC pairs, C objects were on the left. Participants intentionally encoded pairs by creating visual or verbal stories. Pairs were presented in a pseudorandom order, with the constraint that two pairs from the same triad (an AB and its corresponding BC) were not presented in immediate succession. Participants were not made aware of the relationship between AB and BC pairs or that they would be making inference judgments before beginning the experiment.

Following each study phase, participants completed a self-paced three-alternative forced-choice test on all AB and BC pairs (Figure 1B, purple). A cue object (B for AB pairs; C for BC pairs) was presented on the top of the screen with three choice objects on the bottom. Participants were asked to select the choice object that was paired with the cue using a key press. Incorrect options (i.e., foils) were familiar objects from different triads. No feedback was provided. There was a minimum of two trials presented between AB and BC test trials from the same triad; trial order was otherwise random. Participants practiced both study and test tasks before beginning the experiment. Practice pairs were not overlapping, so as not to encourage any strategy in particular before beginning the experiment.

After the training period, participants were told that A and C items were indirectly related through their common association with item B. Instructions were repeated as many times as necessary to ensure understanding. A three-alternative forced-choice test over all inference (AC) associations was administered (Figure 1B, green). C items served as cues, and no feedback was provided.

Analysis of behavioral data. Proportion correct was calculated for each direct pair test (AB/BC) and the inference test (AC). Inference performance was calculated only for AC trials for which the corresponding AB and BC trials were correct in the final direct test. Therefore, any difference in inference performance across groups was not due to an inability to recall the underlying associations. To investigate performance as a function of age group, we performed a 2 × 3 mixed ANOVA with Test (final direct/inference) as the within-subject factor and Age group (child/adolescent/adult) as the between-subject factor. Because of high levels of performance on the final direct pair test, we also computed an average direct pair performance across all four tests for each participant. Average direct pair and inference performance were related to individual differences in structure using multiple regression.

MR Data Acquisition and Analyses

Imaging data were acquired on a 3.0-T Siemens Skyra MRI (Siemens, Erlangen, Germany). Two (or three, if one of the first two images showed motion artifacts via visual inspection) oblique coronal T2-weighted images were acquired perpendicular to the main axis of the HPC (repetition time = 13,150 msec, echo time = 82 msec, 512 × 60 × 512 matrix, 0.4 × 0.4 mm in-plane resolution, 1.5 mm thru-plane resolution, 60 slices, no gap). Coronal images of acceptable quality as determined by visual inspection (see below) were coregistered using ANTS (Avants

et al., 2011) and averaged to boost SNR, yielding a single mean coronal image per participant. Participants for whom the final mean coronal image was of acceptable quality, defined as an absence of motion artifacts that would prevent visualization of the hippocampal sulcus, were included in the analyses. A T1-weighted 3-D MPRAGE volume (256 × 256 × 192 matrix, 1 mm^3 voxels) was also collected.

Hippocampal ROI Definition

HPC ROIs were delineated by hand on each participant's mean coronal image by a single rater (KFG) using established guidelines (Figure 2A; Bonnici et al., 2012; West & Gundersen, 1990). The rater was blind to participant identity, and images were cropped to obscure overall head size.

HPC was segmented into the following subfields: cornu ammonis fields 1 (CA_1) and 2/3 (combined; $CA_{2/3}$), DG, and subiculum. Segmentation was performed across the entire extent of the HPC long axis, with the exception of the most posterior slices on which subfields could not be reliably delineated. For this region, we created a combined posterior HPC ROI. Subfields were summed to create overall HPC ROIs, which were then further segmented into head, body, and tail subregions using anatomical landmarks as follows. Here, we use the term "subregion" to refer to head, body, and tail ROIs that divide the HPC along its long axis, each of which comprises multiple subfields. The posterior boundary of the HPC head was the last slice on which the uncal apex was visible (Poppenk & Moscovitch, 2011; Weiss, Dewitt, Goff, Ditman, & Heckers, 2005). The anterior boundary of the HPC tail was the first slice on which the fornix separated from the HPC (Watson et al., 1992). This process resulted in head, body, and tail ROIs for each participant, as well as CA_1, $CA_{2/3}$, DG, and subiculum subfields within head and body segments. Note that because the HPC tail included a combined posterior HPC ROI for the vast majority of participants, we do not consider subfields within the tail.

Reliability of ROI Definition

We assessed both intra- and interrater reliability to validate our segmentation approach. For the main analyses, all segmentations were performed by a single individual (KFG) who was blinded to participant identity (hereafter, "primary rater"). We further quantified the degree of consistency not only for our primary rater both with herself across a delay (intra-rater reliability) but also with a different rater who was also blinded to participant identity (MLS; hereafter, "secondary rater") referencing the

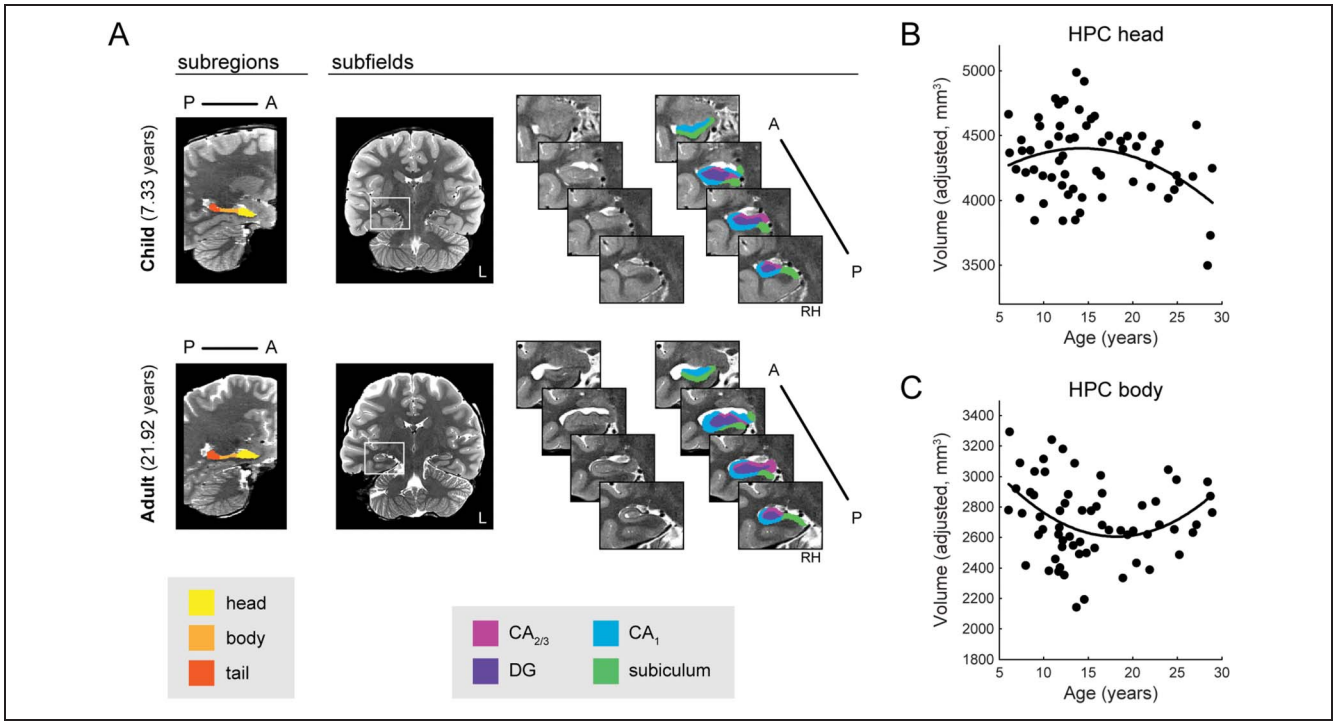

Figure 2. (A) Example ROIs for a representative child (top) and adult (bottom) participant. Left, HPC was divided into head, body, and tail subregions along the anterior–posterior axis. Right, CA_1, $CA_{2/3}$, DG, and subiculum were demarcated across the majority of the anterior–posterior axis. White box on coronal indicates approximate location of inset subfield images. (B) Relationship between age and volume for the HPC head. After excluding multivariate outliers, n = 22 children, 23 adolescents, 19 adults (excluded n = 4). (C) Relationship between age and volume for the HPC body. After excluding multivariate outliers, n = 22 children, 22 adolescents, 19 adults (excluded n = 5). B and C depict adjusted response plots, which show the fitted response as a function of age, averaging out other predictors in the model. Adjusted data were calculated by adding the residual to the fitted value for each point.

same segmentation protocol (interrater reliability). We focused both intra- and interrater reliability analyses on the child group, for whom we reasoned the hippocampi might be the most variable and therefore the most difficult to segment. For the intrarater reliability analysis, the primary rater segmented the children's hippocampi for a second time after a delay of at least 1 year. For the interrater reliability analysis, the secondary rater segmented the children's hippocampi using the same protocol. We computed intraclass correlations (ICC) to measure consistency of averaged measures using a two-way random effects ANOVA model. ICC values near 1 indicate highly consistent segmentation, and values near 0.5 indicate unreliable segmentation. Prior work on hippocampal subfields using the same ICC measure has considered values exceeding 0.70 to represent good consistency (Mueller et al., 2007).

Hippocampal Volume Calculations

Intracranial volume (ICV) was estimated for each participant using Freesurfer (Desikan et al., 2006). We extracted volumes for each ROI and participant. To account for differences in overall head size, volumes for each ROI were adjusted for ICV using an ANCOVA approach (Raz et al., 2005). Specifically, each ROI (including overall HPC) was regressed on ICV across the age range to determine the slope (β_{ICV}) of the relationship between overall head size and ROI volume. Raw ROI volumes were then adjusted to correct for this relationship by subtracting the product of mean-centered ICV measures and β_{ICV} from each ROI. This procedure removes the statistical relationship between ICV and ROI volumes. We then related HPC volumes to (1) age and (2) performance on the statistical learning and associative inference tasks as described below. As recent work suggests that FSL may be superior for estimating ICV in pediatric populations (Sargolzaei et al., 2015), we also verified our findings using ICV estimated using tools available through FSL (following the ENIGMA protocol; enigma.ini.usc.edu). Adjusting for overall head size using the FSL-based ICV estimates did not substantially change any of our findings (results not reported).

Age–Volume Relationships across the HPC Long Axis

We performed multiple linear regression analyses to test for differences in HPC subregion (head, body, tail) volumes across the age range (Table 1). We included ICV-adjusted HPC volume as a nuisance regressor to isolate effects specific to the subregion of interest, rather than changes in the size of the HPC overall. Outliers were excluded for age–volume and volume–performance analyses as described below; see figure captions for the total *n* excluded for each analysis and the final sample size.

First, we identified multivariate outliers for each subregion by applying the box plot rule (Frigge, Hoaglin, & Iglewicz, 1989) to observations with unusually low weights computed using robust regression. Outlier individuals were excluded from the analysis for that subregion. We then tested the following four regression models for each of the three subregions (head, body, and tail; all bilateral), with subregion volume always serving as the dependent variable: (1) a linear main effects model, which included overall HPC volume, Sex, and Age as independent variables; (2) a nonlinear main effects model, which included overall HPC volume, Sex, and Age^2 as independent variables; (3) a linear interaction model, which included overall HPC volume and an Sex × Age interaction as independent variables; and (4) a nonlinear interaction model, which included overall HPC volume as well as Sex × Age and Sex × Age^2 interaction terms as independent variables. We tested models with quadratic terms because prior work has reported a nonlinear relationship between age and HPC volume (Østby et al., 2009). We compared the four models for each subregion using Akaike information criterion corrected for sample size (AIC_c), which penalizes more complex models in the case of small samples (Hurvich & Tsai, 1989). In all models, participants were treated as a random effect. Continuous variables were mean centered for statistical reporting.

Volume–Performance Relationships across the HPC Long Axis

Multiple linear regression analyses were also performed to assess the degree to which HPC subregion volumes (head, body, tail) predicted performance as a function of age, controlling for sex and IQ (Table 2). We chose to control for IQ because we wanted to isolate relationships that were specific to our behavioral measures of interest, rather than related to possible differences in intelligence more generally. Separate models were run for associative inference and statistical learning tasks.

We first identified and excluded multivariate outliers using the robust regression approach described above. We then tested the two possible sets of predictors that we reasoned might best predict variability in performance. On the one hand, if all age groups are relying similarly on HPC structures to perform the tasks, we would expect that to be reflected in a main effect of region size on performance, controlling for the effects of age. Alternatively, if individuals are relying differentially on HPC subregions across the age range, we might instead see an interaction between volume and age. Accordingly, we compared two regression models for each task, with the performance measure of interest (triplet discrimination or inference performance) serving as the dependent variable. For the statistical learning task, the first model included main effects of head, body, and tail volumes predicting performance, with Age, IQ, and Sex

Table 1. All Models Compared for the Volume–Age Analyses

Regression Model	*Adj. R^2*	*Regression Model*	*Adj. R^2*
Subregions		*Follow-up: Subfields within the HPC Head*	
Head		CA_1	
head ~ HPC + Sex + Age	–	CA_1 ~ HPC + Sex + Age	0.76
head ~ HPC + Sex + Age^2	0.78	CA_1 ~ HPC + Sex + Age^2	–
head ~ HPC + Sex × Age	–	CA_1 ~ HPC + Sex × Age	–
head ~ HPC + Sex × Age^2	–	CA_1 ~ HPC + Sex × Age^2	–
Body		$CA_{2/3}$	
body ~ HPC + Sex + Age	–	$CA_{2/3}$ ~ HPC + Sex + Age	0.41
body ~ HPC + Sex + Age^2	0.55	$CA_{2/3}$ ~ HPC + Sex + Age^2	–
body ~ HPC + Sex × Age	–	$CA_{2/3}$ ~ HPC + Sex × Age	–
body ~ HPC + Sex × Age^2	–	$CA_{2/3}$ ~ HPC + Sex × Age^2	–
Tail		DG	
tail ~ HPC + Sex + Age	0.53	DG ~ HPC + Sex + Age	0.49
tail ~ HPC + Sex + Age^2	–	DG ~ HPC + Sex + Age^2	–
tail ~ HPC + Sex × Age	–	DG ~ HPC + Sex × Age	–
tail ~ HPC + Sex × Age^2	–	DG ~ HPC + Sex × Age^2	–
		SUB	
		SUB ~ HPC + Sex + Age	–
		SUB ~ HPC + Sex + Age^2	–
		SUB ~ HPC + Sex × Age	–
		SUB ~ HPC + Sex × Age^2	0.48
Follow-up: Subfields within the HPC Body			
CA_1		DG	
CA_1 ~ HPC + Sex + Age	–	DG ~ HPC + Sex + Age	0.45
CA_1 ~ HPC + Sex + Age^2	0.41	DG ~ HPC + Sex + Age^2	–
CA_1 ~ HPC + Sex × Age	–	DG ~ HPC + Sex × Age	–
CA_1 ~ HPC + Sex × Age^2	–	DG ~ HPC + Sex × Age^2	–
$CA_{2/3}$		SUB	
$CA_{2/3}$ ~ HPC + Sex + Age	0.31	SUB ~ HPC + Sex + Age	0.45
$CA_{2/3}$ ~ HPC + Sex + Age^2	–	SUB ~ HPC + Sex + Age^2	–
$CA_{2/3}$ ~ HPC + Sex × Age	–	SUB ~ HPC + Sex × Age	–
$CA_{2/3}$ ~ HPC + Sex × Age^2	–	SUB ~ HPC + Sex × Age^2	–

Adjusted R^2 is reported for the best-fitting model with the lowest Akaike information criterion corrected for sample size (AIC_C). Region names (e.g., head, HPC) refer to the regional volume.

serving as additional control regressors. The second model for the statistical learning task included all of the predictors listed above as well as all three subregion Volume (head, body, tail) × Age interactions. The models for the associative inference task were identical to those for the statistical learning task, with the exception that average direct performance was also included in both models as an additional predictor of no interest. We then compared the two models for each task using AIC_c. As the primary goal of this study was to relate volume and

age to performance, we focus on only main effects of volume and Volume × Age interactions. In all models, participants were treated as a random effect. Continuous variables were mean centered before statistical analyses.

Exploratory Subfield Models

Models assessing subfield relationships to age (Age–Volume Relationships) and performance (Volume–Performance Relationships) were carried out within subregions for which overall volumes showed a relationship to age or performance, respectively. Multiple linear regression analyses were conducted as described above, with the subfield volumes (CA_1, $CA_{2/3}$, DG, subiculum) serving as the measures of interest. More specifically, for the age–volume relationships, four models were tested for each subfield as described above, with each subfield volume serving as the dependent variable in turn. For the volume–performance relationships, the two regression models included the main effects of volume for the four subfields, or the main effects plus Volume × Age interactions for all four subfields simultaneously. As in the subregion models, we then selected the best model for each analysis by comparing AIC_c. For this analysis, we interrogated subfield volumes restricted to the subregions (i.e., head, body, or tail) that showed significant relationships to age and/or performance. We provide these results for completeness; however, we view them as exploratory in nature given current disagreement about the ability to segment the HPC head into subfields using MRI. Readers should weigh this caveat seriously when interpreting the subfield findings.

Relationships between Tasks

Because we found that both statistical learning and associative inference task performance improved with age, one possibility is that the two tasks are redundant measures of a single mnemonic function and a common neural substrate. Alternatively, performance on each task might be associated with unique variance in behavioral and/or structural development. To assess these possibilities, we first tested for a relationship between performances on the two tasks using multiple regression, controlling for the effects of age.

Table 2. All Models Compared for the Volume–Performance Analyses

Regression Model	*Adj. R^2*
Subregions	
Statistical learning	
discrimination ~ head + body + tail + age + IQ + sex	0.39
→ discrimination ~ head + body + tail + age + IQ + sex + inference	0.46
discrimination ~ head × Age + body × Age + tail × Age + IQ + sex	–
Associative inference	
inference ~ head + body + tail + age + IQ + sex + direct	0.76
→ inference ~ head + body + tail + age + IQ + sex + direct + discrimination	0.71
inference ~ head × Age + body × Age + tail × Age + IQ + sex + direct	–
Follow-up: Subfields within the HPC Head	
Statistical learning	
discrimination ~ CA_1 + $CA_{2/3}$ + DG + SUB + age + IQ + sex	0.48
→ discrimination ~ CA_1 + $CA_{2/3}$ + DG + SUB + age + IQ + sex + inference	0.63
discrimination ~ CA_1 × Age + $CA_{2/3}$ × Age + DG × Age + SUB × age + IQ + sex	–
Associative inference	
inference ~ CA_1 + $CA_{2/3}$ + DG + SUB + age + IQ + sex + direct	–
inference ~ CA_1 × Age + $CA_{2/3}$ × Age + DG × Age + SUB × Age + IQ + sex + direct	0.74
→ inference ~ CA_1 × Age + $CA_{2/3}$ × Age + DG × Age + SUB × Age + IQ + sex + direct + discrimination	0.75

Adjusted R^2 is reported for the best-fitting model with the lowest Akaike information criterion corrected for sample size (AIC_C). Discrimination, triplet discrimination performance on statistical learning task. Inference, inference performance on associative inference task. Direct, average direct memory performance across the four study repetitions on the associative inference task. Region names (e.g., head, CA_1) refer to the regional volume.

We next performed additional analyses to determine whether performance differences on the two tasks could be explained by the same variability in HPC volumes. To do this, we added performance on the alternate task to each of the best-fitting models determined through the model selection approach described previously. That is, we added inference performance to the model predicting triplet discrimination performance for the statistical learning task, and vice versa (Table 2). This approach allowed us to investigate the relationship between HPC volumes and performance on one task, controlling for variance explained by the other. If HPC volumes no longer predict performance in these models, it would suggest that the two tasks relate to common variability in structure. However, if HPC volumes remain significant predictors of performance, it would indicate that the two tasks each are each associated with unique variance.

To ensure that multicollinearity (e.g., among performance on the two tasks) was not adversely impacting our regression results, we performed collinearity diagnostics on all volume–performance models. We found variance inflation factors (VIFs) to be within the acceptable range in all cases (all subregion model VIF < 1.95 and all exploratory subfield model VIF < 4.13, where VIF < 10 is typically deemed acceptable).

RESULTS

Reliability of HPC Segmentation

We first assessed intrarater reliability, which compared segmentations of the child group performed by the primary rater on two separate occasions separated by a delay of at least 1 year. Reliability was excellent in defining overall HPC (ICC = 0.92), as well in delineating the hippocampal head (ICC = 0.95) and body (ICC = 0.96). Delineation of the hippocampal tail was less reliable (ICC = 0.67). We also found high consistency for subfields within the hippocampal head (CA_1: 0.89, $CA_{2/3}$: 0.76, subiculum: 0.84, DG: 0.91). Consistency was also generally high within the hippocampal body (CA_1: 0.92, subiculum: 0.91, DG = 0.90), with the lowest reliability observed in $CA_{2/3}$ (ICC = 0.61).

Next, we assessed consistency between the primary and secondary raters. Results revealed excellent reliability in defining the overall HPC (ICC = 0.96) as well as in identifying head (ICC = 0.96) and body (ICC = 0.83) subregions. Again, identification of the tail subregion was less reliable (ICC = 0.52). Within the head of the HPC, we found generally high consistency between the primary and secondary raters in identifying subfields (CA_1: 0.93, $CA_{2/3}$: 0.88, subiculum: 0.76, DG: 0.90). Notably, the lowest consistency within the head was observed in subiculum, which converges with prior reports (Yushkevich et al., 2010; Mueller et al., 2007). Subfield delineation within the hippocampal body was also generally consistent across raters (CA_1: 0.78, $CA_{2/3}$: 0.76, subiculum: 0.88; DG: 0.70). Overall, our reliability results are in the same range as previous reports for manual demarcation of hippocampal subfields (Lee et al., 2014; Wisse et al., 2012; Yushkevich et al., 2010; Mueller et al., 2007). Moreover, reliability measures were approximately equal when comparing subfields delineated within the HPC head to those defined within the body for both intra- and interrater analyses. However, across both intra- and interrater reliability analyses, consistency in defining the HPC tail was relatively poor.

Differential Development across HPC Axis

Our first goal was to determine how HPC structure changes from middle childhood through adulthood, testing specifically whether different HPC longitudinal subregions (Figure 2A) show different developmental patterns. We interrogated the relationship between age and volumes for HPC head, body, and tail subregions by first selecting the best among four possible multiple regression models. We then focus specifically on the age-related effects, which index the relationship between age and volume controlling for other predictors in the model (overall HPC volume, sex). Within those regions showing significant age–volume relationships, we then interrogate subfields as an exploratory analysis investigating potential differences within the overall head, body, and tail subregions.

For the hippocampal head, the best fitting model included a significant quadratic effect of age on volume (head volume ~ HPC volume + sex + age^2; overall model fit: adjusted $R^2 = 0.78$, $F(4, 59) = 56.5$, $p < 1 \times 10^{-18}$; reliability of age^2 effect: $\beta = -0.14$, $p = .047$; all statistics reflect standardized β), with volumes increasing through adolescence and decreasing into adulthood (Figure 2B). Differences in the overall volume of the HPC body over age were also best described by a model including a quadratic effect of age (body volume ~ HPC volume + sex + age^2; adjusted $R^2 = 0.55$, $F(4, 58) = 19.6$, $p < 1 \times 10^{-9}$), with both the main ($\beta = -0.22$, $p = .03$) and quadratic ($\beta = 0.30$, $p < .01$) effects of Age reaching statistical significance. This region showed decreases through adolescence followed by volume increases into adulthood (Figure 2C), contrasting with the relationship observed in the HPC head. Individual differences in hippocampal tail volumes were best accounted for by a model including the main effect of Age (tail volume ~ HPC volume + sex + age; adjusted $R^2 = 0.53$, $F(3, 58) = 24.1$, $p < 1 \times 10^{-9}$), although the effect of Age was not statistically significant ($\beta = 0.14$, $p = .11$).

Exploratory Subfield Results

Because there is controversy surrounding the ability to segment HPC head into subfields, all results relating to subfields within the head should be interpreted with

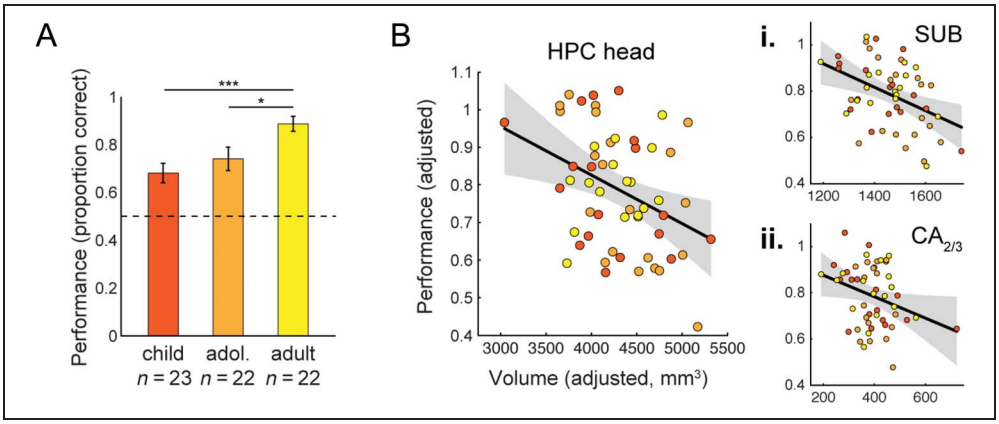

Figure 3. Statistical learning behavior and volume–performance relationships. (A) Behavioral performance on triplet discrimination task by age group. There was a significant effect of age group ($p < .01$; not marked on chart). Bar heights represent group means; error bars denote *SEM*. Asterisks indicate significant two-sample t tests at $*p < .05$ and $***p < .001$. (B) Negative relationship between HPC head volume and triplet discrimination performance. After excluding multivariate outliers, $n = 18$ children, 21 adolescents, 17 adults (excluded $n = 4$). Follow-up subfield analysis showed significant negative relationship for subiculum (Bi) and $CA_{2/3}$ (Bii). After excluding multivariate outliers, $n = 17$ children, 19 adolescents, 17 adults for both subfield plots (excluded $n = 7$). For B, data are presented as individual coefficient plots to depict the main effect of volume on triplet discrimination performance, controlling for all other predictors. Shaded regions represent 95% confidence intervals.

caution. With that caveat in mind, we found an association between age and the volume of CA_1 within the HPC head. For this region, the best-fitting regression model (CA_1 volume ~ HPC volume + sex + age; adjusted $R^2 = 0.76$, $F(3, 60) = 67.9$, $p < 1 \times 10^{-18}$) showed a significant negative relationship between Age and CA_1 subfield volume ($\beta = -0.21$, $p < .01$). No other subfield in the head showed a significant relationship with age (p for all age-related effects $> .06$).

Within the body, subiculum and DG volumes were significantly related to age. DG volume was positively related to Age (DG volume ~ HPC volume + sex + age; adjusted $R^2 = 0.45$, $F(3, 58) = 17.5$, $p < 1 \times 10^{-7}$, $\beta = 0.26$, $p < .01$). Subiculum volume, in contrast, showed a significant negative relationship with age (subiculum volume ~ HPC volume + sex + age; adjusted $R^2 = 0.45$, $F(3, 58) = 17.6$, $p < 1 \times 10^{-7}$, $\beta = -0.31$, $p < .01$). The relationships between age and volume were not significant for CA_1 or $CA_{2/3}$ (both $p > .08$).

Statistical Learning and Inference Performance Improves over Development

Performance improved across the age range for both statistical learning and associative inference tasks. In the statistical learning task, all age groups performed above chance on the triplet discrimination test (all $p < 1 \times 10^{-3}$). Triplet discrimination performance (Figure 3A) differed significantly across age groups ($F(2, 64) = 6.64$, $p < .01$), with adults performing significantly better than both children ($t(43) = 3.96$, $p < .001$) and adolescents ($t(42) = 2.52$, $p = .02$). There was no performance difference between child and adolescent groups ($t(43) = 0.93$, $p = .36$).

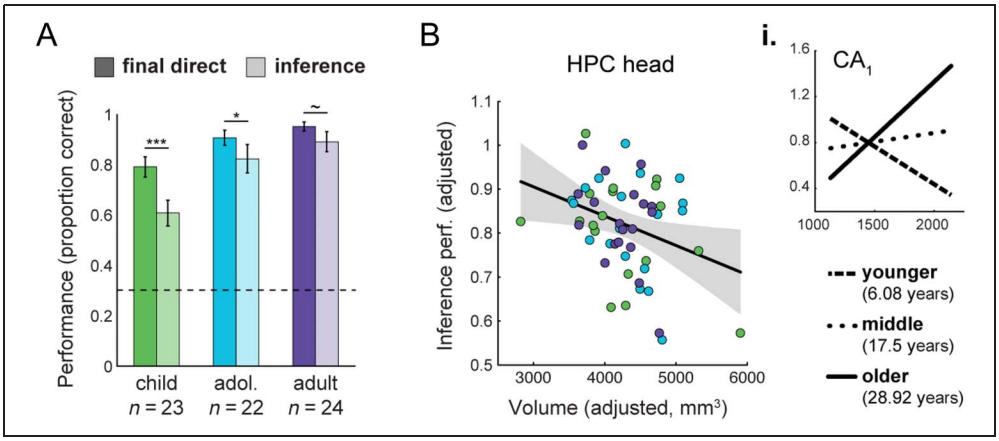

Figure 4. Associative inference behavior and volume–performance relationships. (A) Behavioral performance on final direct and inference tests by age group. There was a significant Age group × Test trial type interaction ($p < .05$), as well as main effects of Test trial type ($p < 1 \times 10^{-5}$) and Age group ($p < .001$). Bar heights represent group means; error bars denote *SEM*. Asterisks indicate significant paired t tests at $*p < .05$ and $***p < .001$; statistical trend indicated with $\sim p < .1$. (B) Main effect of HPC head volume on inference performance. Individual coefficient plot as depicted in Figure 3B. After excluding multivariate outliers, $n = 18$ children, 19 adolescents, 19 adults (excluded $n = 6$). (Bi) Follow-up subfield analysis within the HPC head revealed a significant Volume × Age interaction for CA_1. Plot shows interaction effect as predicted changes in inference performance (y axis) across a range of volumes (x axis) for three fixed values (lines) of age. Fixed values were chosen to produce a large effect in inference performance, thus allowing for easy visualization of the interaction. After excluding multivariate outliers, $n = 18$ children, 21 adolescents, 19 adults (excluded $n = 4$).

For the associative inference task (Figure 4A), all groups performed above chance for both final direct and inference tests (all $p < .001$). There were main effects of test (final direct/inference; $F(1, 66) = 25.97$, $p < 1 \times 10^{-5}$) and age group ($F(2, 66) = 10.21$, $p < .001$), as well as a Test × Age group interaction ($F(2, 66) = 3.15$, $p < .05$). Follow-up pairwise comparisons revealed lower inference than final direct performance in child ($t(22) = 4.26$, $p < .001$) and adolescent ($t(21) = 2.47$, $p = .02$) groups (*ns* for adults: $t(23) = 1.81$, $p = .08$). Notably, because inference performance was examined only when the underlying direct pairs were remembered correctly, this difference cannot be attributed to differences in baseline memory ability. Children performed worse overall than both adolescents (final direct: $t(43) = 2.31$, $p = .03$; inference: $t(43) = 2.85$, $p < .01$) and adults (final direct: $t(45) = 3.72$, $p < .001$; inference: $t(45) = 4.40$, $p < .001$); there were no performance differences between adolescents and adults (both $|t(44)| < 1.30$, $p > .20$). Inference performance was significantly related to triplet discrimination performance controlling for age ($\beta = 0.47$, $p < .001$), suggesting that a common associative binding process supports performance on both tasks.

HPC Head Volumes Predict Statistical Learning and Inference Performance

The above results demonstrate differences in both HPC volume and performance across the age range. Our next goal was to link these two measures by investigating the relationship between HPC subregion volumes and performance on both cognitive tasks.

Statistical Learning Task

For the statistical learning task, the best-fitting model was the main effects model (performance ~ HPC head + HPC body + HPC tail + age + IQ + sex; adjusted $R^2 = 0.39$, $F(6, 49) = 6.81$, $p < 1 \times 10^{-4}$), and the coefficient on HPC head volume was significant ($\beta = -0.33$, $p < .01$; Figure 3B). There were no significant effects of HPC body or tail volume (both $p > .11$). This result suggests that, after accounting for age, individual differences in HPC head volume show a negative relationship to triplet discrimination performance. The main effect of HPC head volume remained significant after controlling for inference performance ($\beta = -0.36$, $p < .01$).

Given the age–volume relationship we observed in the HPC head (Figure 3B), we explored whether this volume–performance relationship differed across subfields within the HPC head. The main effects model best described the data (performance ~ CA_1 + $CA_{2/3}$ + DG + subiculum + age + IQ + sex; adjusted $R^2 = 0.48$, $F(7, 45) = 7.80$, $p < 1 \times 10^{-5}$). Within the HPC head, individual differences in both subiculum and $CA_{2/3}$ volumes were negatively related to triplet discrimination performance, controlling for age (coefficient on volume for subiculum: $\beta = -0.43$, $p < .01$, Figure 3Bi; and coefficient on volume for $CA_{2/3}$: $\beta = -0.34$, $p = .04$, Figure 3Bii). Coefficients on CA_1 and DG were not significant (both $p > .11$). The relationship between subiculum volume and triplet discrimination performance remained significant when inference performance was added to the model (subiculum: $\beta = -0.49$, $p < .001$; $CA_{2/3}$: $\beta = -0.34$, $p = .02$). As with the main volume–age findings, these subfield results should be considered carefully given the controversy surrounding identification of subfields within the HPC head using MRI.

Associative Inference Task

For the associative inference task, the best-fitting regression model (performance ~ HPC head + HPC body + HPC tail + direct performance + age + IQ + sex) included a significant effect of HPC head volume on AC performance. As this model controls for individual differences in memory for the direct pairs, any relationship between volume and inference performance is robust to any possible covariance among these factors. The model fit was significant (adjusted $R^2 = 0.76$, $F(7, 48) = 26.1$, $p = 1 \times 10^{-13}$), with a negative relationship between HPC head volume and inference performance ($\beta = -0.18$, $p = .02$; Figure 4B). Coefficients on HPC body and tail volume were not significant (both $p > .42$). The main effect of HPC head volume remained significant after controlling for triplet discrimination performance ($\beta = -0.19$, $p = .05$).

The best-fitting model for the follow-up subfield analysis within the HPC head included Volume × Age interactions (performance ~ CA_1 × age + $CA_{2/3}$ × age + DG × age + subiculum × age + direct performance + IQ + sex; adjusted $R^2 = 0.74$, $F(12, 45) = 14.5$, $p < 1 \times 10^{-10}$). The main effects of Volume are not of interest in this case, as they would reflect the relationship between volume and performance at a particular age; thus, we focus here on the interactions. CA_1 volume was the only subfield to show a significant interaction with age ($\beta = 0.33$, $p = .01$, Figure 4Bi; for all other subfields, $p > .19$), with the volume–inference relationship increasing across the age range from negative in younger to positive in older participants. The CA_1 × Age interaction remained significant after triplet discrimination performance was added to the model ($\beta = 0.38$, $p < .01$). Again, these CA_1 findings should be interpreted with caution given the caveats noted above.

DISCUSSION

In this study, we characterized the link between differences in HPC volumes and extraction of associative structure across children, adolescents, and adults. Prior work suggests that maturation of HPC binding may give rise to behavioral gains in memory for within-event associations

over development. Our findings further suggest that there are also developmental changes in cross-episode binding ability. We found evidence for subtle differences in HPC structure over the age range, with dissociable developmental patterns across anterior (head) and more posterior (body) regions. These structural differences were coupled with gains in memory performance on both statistical learning and associative inference tasks, suggesting that the ability to extract structure across multiple episodes improves from childhood to adulthood.

Dissociable Structural Changes across HPC Anterior–Posterior Axis

In contrast to the majority of prior work on hippocampal development (Krogsrud et al., 2014; Tamnes et al., 2014), we were interested in characterizing structural differences as a function of position along the anterior–posterior HPC axis. We found a nonlinear relationship between age and HPC head volume, with volume increases early in the age range followed by decreases into adulthood. Interestingly, this pattern parallels predictions from a computational account of neural development, which shows that optimal memory performance arises when early synaptic overgrowth is followed by selective pruning (Chechik, Meilijson, & Ruppin, 1998; see also Eckenhoff & Rakic, 1991). In contrast, the HPC body showed a U-shaped trajectory, with early decreases followed by increases in the adult range. Broadly, the dissociable patterns of development we observed across the HPC head and body are consistent with prior work, which has demonstrated volume decreases in anterior and increases in posterior HPC regions across development (Demaster et al., 2013; Gogtay et al., 2006). Notably, two previous studies that have concluded a linear decline in HPC head volume over development used either a smaller age range (e.g., 8–18 years; Daugherty, Yu, et al., 2015) or directly compared only two narrow child and adult groups (Demaster et al., 2013). However, our data suggest that it may be necessary to investigate development across a wider age range (here, 6–30 years) to uncover nonlinear patterns, with sufficient sampling at the youngest ages and in adolescence being particularly critical. Because of the cross-sectional nature of our study, we cannot definitively conclude that these observed differences would mirror within-individual changes, so future work using longitudinal paradigms will provide further insight.

As additional exploratory analyses, we investigated the development of hippocampal subfields. Relating our results to the existing developmental literature is a challenge for several reasons. First, the bulk of existing research has characterized anatomical differences across age, without linking the anatomy to behavior (for exceptions, see Lee et al., 2014; Tamnes et al., 2014; Demaster et al., 2013). Thus, the behavioral relevance of developmental differences remains largely unknown. Second, the size of different HPC subfields are most often measured either across the entire anterior–posterior extent of HPC (Krogsrud et al., 2014; Tamnes et al., 2014) or only within the HPC body (Daugherty, Bender, Raz, & Ofen, 2016; Lee et al., 2014). These differences across studies have yielded mixed findings in the literature, and it remains a challenge to provide a solid characterization of HPC development. Third, controversy remains as to how to best segment the HPC into subfields, with some researchers arguing against dividing the head into subfields at all. Although our reliability in delineating subfields within the head was on par with published standards for HPC body segmentation, the degree to which this tracing protocol captures the true structure remains unknown given the paucity of histological data (particularly in pediatric samples). We thus encourage caution in interpreting our subfield findings, especially within the head, and acknowledge that future work will be needed to further understand HPC subfield development. Importantly, however, we underscore that the controversy described above is unrelated to our main findings, which were at the subregion (head, body, tail) level. Our results thus add to a growing body of work suggesting that HPC structural development is subtle, varying as a function of both subfield and position along the long axis. It will be important in future studies to consider both of these factors, as well as structure–behavior relationships, to better characterize the development of this system.

Performance Increases across Development

Lesion studies have shown that behavior in both the statistical learning and associative inference tasks critically depends on HPC (Schapiro et al., 2014; DeVito, Kanter, & Eichenbaum, 2010). We found that performance on both tasks increased across the age range, suggesting gains in HPC function into adulthood. These findings are consistent with prior behavioral work suggesting that, although even young children are capable of inference (Bauer, Varga, King, Nolen, & White, 2015; Andrews & Halford, 1998; Halford, 1984), this ability continues to improve through at least late childhood (Townsend, Richmond, Vogel-Farley, & Thomas, 2010). Performance on the associative inference and statistical learning tasks was also related even when controlling for age, suggesting a common memory mechanism supporting the ability to extract associative structure across these two tasks.

HPC Head Volume Predicts Performance

Across the age range, we found negative relationships between HPC head volume and behavioral performance, controlling for age, IQ, sex (all for both tasks), and average direct performance (for the associative inference task only). Although the fact that this effect was unrelated to

age might appear counterintuitive, one possibility is that the repeated learning exposures in both tasks promoted HPC-based encoding in all ages; it has been proposed that children are capable of engaging HPC encoding mechanisms earlier in development when experiences are repeated multiple times relative to when they are seen just once (Gómez & Edgin, 2015; Lavenex & Banta Lavenex, 2013). Previous studies across child, adolescent, and young adult samples have reported similar negative correlations between overall HPC volumes and memory performance (Yurgelun-Todd et al., 2003; Chantôme et al., 1999; Foster et al., 1999).

There are at least two possible explanations for the sign of this relationship. One possibility (Van Petten, 2004) is that the rest of the brain gets larger (e.g., white matter volume increases; Sowell, Trauner, Gamst, & Jernigan, 2002; Giedd et al., 1999; Reiss, Abrams, Singer, Ross, & Denckla, 1996), whereas HPC itself remains unchanged. This pattern could give rise to the apparent negative relationships between HPC volume and performance, perhaps resulting from a positive association between behavior and extra-HPC volume. However, this possibility does not readily account for subregional differences, with the opposite patterns in more posterior HPC regions (Demaster et al., 2013; Poppenk & Moscovitch, 2011) being particularly problematic to explain under this hypothesis. Alternatively, volume decreases in the HPC head and increases in the HPC body between adolescents and adults may reflect pruning and proliferation, respectively (Foster et al., 1999). Under this framework, insufficient pruning in the HPC head may be associated with worse performance, as pruning of extraneous connections (Cowan, Fawcett, Leary, & Stanfield, 1984) may increase processing efficiency (Chechik et al., 1998). Pruning within the HPC head would yield higher representational overlap for related experiences, whereas proliferation in more posterior regions of HPC would allow for distinct, nonoverlapping traces. We present these interpretations with the caveat that the MRI methodologies we employ here—though generally thought to capture developmental changes such as axonal myelination, dendritic arborization, and synaptic pruning (Huttenlocher, 1990)—are unable to quantify such microstructure directly. Thus, the precise mechanisms giving rise to the results presented here remain an open question.

As supplementary analyses, we also investigated how subfield volumes within the HPC head were related to behavior. Our data revealed a negative relationship between volumes of both $CA_{2/3}$ and subiculum and statistical learning performance after controlling for the effects of age. Prior fMRI work using a similar temporal association task has suggested a role for CA_3 and possibly subiculum in predicting upcoming items (Schapiro et al., 2012). One possible interpretation of the present result is that CA_3 reinstates upcoming items in the sequence via pattern completion (Leutgeb & Leutgeb, 2007; Norman & O'Reilly, 2003) and projects them to the subiculum; however, the role of subiculum in prediction and retrieval remains poorly understood (Ketz, Morkonda, & O'Reilly, 2013).

Subfield analyses for associative inference revealed a significant interaction of CA_1 volume and age, with a negative volume–performance relationship in younger participants becoming positive in older participants. Previous work has suggested that CA_1 detects mismatches when current experience deviates from memory (O'Reilly & Rudy, 2001), which is thought to promote new encoding (Schlichting & Preston, 2015; Duncan, Ketz, Inati, & Davachi, 2012; Shohamy & Wagner, 2008). Moreover, a high-resolution fMRI study in adults using a similar task has demonstrated engagement of the CA_1 subfield during learning (Schlichting et al., 2014), consistent with the idea that processing in this region promotes integration across episodes. Although speculative, one possible interpretation of the present finding is that, relative to children, adults more readily extract cross-episode associations from the environment, laying down flexible memory traces that can be applied in future scenarios.

Conclusions

Our results support the notion that episodic memory is not fully mature early in childhood (Ghetti & Bunge, 2012)—rather, the HPC continues to develop throughout adolescence, both in terms of structure and function. We showed behavioral gains into adulthood, consistent with the idea that cross-episode binding improves throughout adolescence as the underlying structure changes.

Acknowledgments

The authors thank Jessica Church-Lang, Michael Mack, Tammy Tran, and Amelia Wattenberger for assistance with participant recruitment, data collection, and helpful discussions. This work was supported by the National Institutes of Health (grants R01MH100121 and R21HD083785 to A. R. P. and R01EY021755 to N. B. T.-B.), the National Science Foundation CAREER Award (grants 1056019 to A. R. P.), a University of Texas Research Grant to A. R. P., and the Department of Defense through the National Defense Science and Engineering Graduate Fellowship (NDSEG) Program (M. L. S.).

Reprint requests should be sent to Alison R. Preston, Center for Learning and Memory, The University of Texas at Austin, 1 University Station, C7000, Austin, TX 78712, or via e-mail: apreston@utexas.edu.

REFERENCES

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4-18 and 1991 profile*. Burlington, VT: Department of Psychiatry, University of Vermont.

Andrews, G., & Halford, G. S. (1998). Children's ability to make transitive inferences: The importance of premise integration and structural complexity. *Cognitive Development, 13,* 479–513.

Aslin, R. N., & Newport, E. L. (2012). Statistical learning: From acquiring specific items to forming general rules. *Current Directions in Psychological Science, 21,* 170–176.
Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., & Gee, J. C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage, 54,* 2033–2044.
Bauer, P. J., Varga, N. L., King, J. E., Nolen, A. M., & White, E. A. (2015). Semantic elaboration through integration: Hints both facilitate and inform the process. *Journal of Cognition and Development, 16,* 351–369.
Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition, 50,* 7–15.
Billingsley, R. L., Smith, M. L., & McAndrews, M. P. (2002). Developmental patterns in priming and familiarity in explicit recollection. *Journal of Experimental Child Psychology, 82,* 251–277.
Bonnici, H. M., Chadwick, M. J., Kumaran, D., Hassabis, D., Weiskopf, N., & Maguire, E. A. (2012). Multi-voxel pattern analysis in human hippocampal subfields. *Frontiers in Human Neuroscience, 6,* 1–13.
Brainerd, C. J., Holliday, R. E., & Reyna, V. F. (2004). Behavioral measurement of remembering phenomenologies: So simple a child can do it. *Child Development, 75,* 505–522.
Carlson, S. M., & Moses, L. J. (2013). Individual differences in inhibitory control and children's theory of mind. *Child Development, 72,* 1032–1053.
Chantôme, M., Perruchet, P., Hasboun, D., Dormont, D., Sahel, M., Sourour, N., et al. (1999). Is there a negative correlation between explicit memory and hippocampal volume? *Neuroimage, 10,* 589–595.
Chechik, G., Meilijson, I., & Ruppin, E. (1998). Synaptic pruning in development: A computational account. *Neural Computation, 10,* 1759–1777.
Collin, S. H. P., Milivojevic, B., & Doeller, C. F. (2015). Memory hierarchies map onto the hippocampal long axis in humans. *Nature Neuroscience, 18,* 1–5.
Cowan, W. M., Fawcett, J. W., Leary, D. D. M. O., & Stanfield, B. B. (1984). Regressive events in neurogenesis. *Science, 225,* 1258–1265.
Crone, E. A., Wendelken, C., van Leijenhorst, L., Honomichl, R. D., Christoff, K., & Bunge, S. A. (2009). Neurocognitive development of relational reasoning. *Developmental Science, 12,* 55–66.
Daugherty, A. M., Bender, A. R., Raz, N., & Ofen, N. (2016). Age differences in hippocampal subfield volumes from childhood to late adulthood. *Hippocampus, 26,* 220–228.
Daugherty, A. M., Yu, Q., Flinn, R., & Ofen, N. (2015). A reliable and valid method for manual demarcation of hippocampal head, body, and tail. *International Journal of Developmental Neuroscience, 41,* 115–122.
Demaster, D. M., Pathman, T., Lee, J. K., & Ghetti, S. (2013). Structural development of the hippocampus and episodic memory: Developmental differences along the anterior/posterior axis. *Cerebral Cortex, 24,* 3036–3045.
Derogatis, L. R. (1977). *SCL-90-R: Administration, scoring and procedures: Manual 1*. Baltimore, MD: Clinical Psychometric Research.
Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage, 31,* 968–980.
DeVito, L. M., Kanter, B. R., & Eichenbaum, H. (2010). The hippocampus contributes to memory expression during transitive inference in mice. *Hippocampus, 20,* 208–217.
Duncan, K., Ketz, N., Inati, S. J., & Davachi, L. (2012). Evidence for area CA1 as a match/mismatch detector: A high-resolution fMRI study of the human hippocampus. *Hippocampus, 22,* 389–398.
Eckenhoff, M. F., & Rakic, P. (1991). A quantitative analysis of synaptogenesis in the molecular layer of the dentate gyrus in the rhesus monkey. *Brain Research. Developmental Brain Research, 64,* 129–135.
Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12,* 499–504.
Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 28,* 458–467.
Foster, J. K., Meikle, A., Goodson, G., Mayes, A. R., Howard, M., Sünram, S. I., et al. (1999). The hippocampus and delayed recall: Bigger is not necessarily better? *Memory, 7,* 715–733.
Frigge, M., Hoaglin, D. C., & Iglewicz, B. (1989). Some implementations of the boxplot. *The American Statistician, 43,* 50–54.
Frost, R., Siegelman, N., Narkiss, A., & Afek, L. (2013). What predicts successful literacy acquisition in a second language? *Psychological Science, 24,* 1243–1252.
Ghetti, S., & Bunge, S. A. (2012). Neural changes underlying the development of episodic memory during middle childhood. *Developmental Cognitive Neuroscience, 2,* 381–395.
Giedd, J. N., Blumenthal, J., Jeffries, N. O., Castellanos, F. X., Liu, H., Zijdenbos, A., et al. (1999). Brain development during childhood and adolescence: A longitudinal MRI study. *Nature Neuroscience, 2,* 861–863.
Giedd, J. N., Vaituzis, A., Hamburger, S. D., Lange, N., Rajapakse, J. C., Kaysen, D., et al. (1996). Quantitative MRI of the temporal lobe, amygdala, and hippocampus in normal human development: Ages 4–18 years. *Journal of Comparative Neurology, 366,* 223–230.
Gogtay, N., Nugent, T. F., Herman, D. H., Ordonez, A., Greenstein, D. K., Hayashi, K. M., et al. (2006). Dynamic mapping of normal human hippocampal development. *Hippocampus, 16,* 664–672.
Gómez, R. L., & Edgin, J. O. (2015). The extended trajectory of hippocampal development: Implications for early memory development and disorder. *Developmental Cognitive Neuroscience, 18,* 57–69.
Halford, G. S. (1984). Can young children integrate premises and serial order tasks? *Transitivity, 93,* 65–93.
Hsu, N. S., Schlichting, M. L., & Thompson-Schill, S. L. (2014). Feature diagnosticity affects representations of novel and familiar objects. *Journal of Cognitive Neuroscience, 26,* 2735–2749.
Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometricka, 76,* 297–307.
Huttenlocher, P. R. (1990). Morphometric study of human cerebral cortex development. *Neuropsychologia, 28,* 517–527.
Ketz, N., Morkonda, S. G., & O'Reilly, R. C. (2013). Theta coordinated error-driven learning in the hippocampus. *PLoS Computational Biology, 9,* e1003067.
Krogsrud, S. K., Tamnes, C. K., Fjell, A. M., Amlien, I., Grydeland, H., Sulutvedt, U., et al. (2014). Development of hippocampal subfield volumes from 4 to 22 years. *Human Brain Mapping, 5657,* 5646–5657.
Lavenex, P., & Banta Lavenex, P. (2013). Building hippocampal circuits to learn and remember: Insights into the development of human memory. *Behavioural 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,* 162–171.
Leutgeb, S., & Leutgeb, J. K. (2007). Pattern separation, pattern completion, and new neuronal codes within a continuous CA3 map. *Learning & Memory (Cold Spring Harbor, N.Y.), 14,* 745–757.
Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S. J., et al. (2000). Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences, U.S.A., 97,* 4398–4403.
Maguire, E. A., Woollett, K., & Spiers, H. J. (2006). London taxi drivers and bus drivers: A structural MRI and neuropsychological analysis. *Hippocampus, 16,* 1091–1101.
Mueller, S. G., Stables, L., Du, A. T., Schuff, N., Truran, D., Cashdollar, N., et al. (2007). Measurement of hippocampal subfields and age-related changes with high resolution MRI at 4T. *Neurobiology of Aging, 28,* 719–726.
Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review, 110,* 611–646.
O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review, 108,* 311–345.
Østby, Y., Tamnes, C. K., Fjell, A. M., Westlye, L. T., Due-Tønnessen, P., & Walhovd, K. B. (2009). Heterogeneity in subcortical brain development: A structural magnetic resonance imaging study of brain maturation from 8 to 30 years. *Journal of Neuroscience, 29,* 11772–11782.
Paller, K. A., & Wagner, A. D. (2002). Observing the transformation of experience into memory. *Trends in Cognitive Sciences, 6,* 93–102.
Poppenk, J., & Moscovitch, M. (2011). A hippocampal marker of recollection memory ability among healthy young adults: Contributions of posterior and anterior segments. *Neuron, 72,* 931–937.
Preston, A. R., Shrager, Y., Dudukovic, N., & Gabrieli, J. D. E. (2004). Hippocampal contribution to the novel use of relational information in declarative memory. *Hippocampus, 14,* 148–152.
Raz, N., Lindenberger, U., Rodrigue, K. M., Kennedy, K. M., Head, D., Williamson, A., et al. (2005). Regional brain changes in aging healthy adults: General trends, individual differences and modifiers. *Cerebral Cortex, 15,* 1676–1689.
Reiss, A. L., Abrams, M. T., Singer, H. S., Ross, J. L., & Denckla, M. B. (1996). Brain development, gender and IQ in children. A volumetric imaging study. *Brain, 119,* 1763–1774.
Sargolzaei, S., Sargolzaei, A., Cabrerizo, M., Chen, G., Goryawala, M., Pinzon-Ardila, A., et al. (2015). Estimating intracranial volume in brain research: An evaluation of methods. *Neuroinformatics, i,* 427–441.
Schacter, D. L., & Wagner, A. D. (1999). Medial temporal lobe activations in fMRI and PET studies of episodic encoding and retrieval. *Hippocampus, 9,* 7–24.
Schapiro, A. C., Gregory, E., Landau, B., McCloskey, M., & Turk-Browne, N. B. (2014). The necessity of the medial temporal lobe for statistical learning. *Journal of Cognitive Neuroscience, 26,* 1736–1747.
Schapiro, A. C., Kustner, L. V., & Turk-Browne, N. B. (2012). Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Current Biology, 22,* 1622–1627.
Schlichting, M. L., Mumford, J. A., & Preston, A. R. (2015). Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nature Communications, 6,* 8151.
Schlichting, M. L., & Preston, A. R. (2015). Memory integration: Neural mechanisms and implications for behavior. *Current Opinion in Behavioral Sciences, 1,* 1–8.
Schlichting, M. L., Zeithamova, D., & Preston, A. R. (2014). CA1 subfield contributions to memory integration and inference. *Hippocampus, 24,* 1248–1260.
Shohamy, D., & Wagner, A. D. (2008). Integrating memories in the human brain: Hippocampal-midbrain encoding of overlapping events. *Neuron, 60,* 378–389.
Sowell, E. R., Trauner, D. A., Gamst, A., & Jernigan, T. L. (2002). Development of cortical and subcortical brain structures in childhood and adolescence: A structural MRI study. *Developmental Medicine and Child Neurology, 44,* 4–16.
Strange, B. A., Witter, M. P., Lein, E. S., & Moser, E. I. (2014). Functional organization of the hippocampal longitudinal axis. *Nature Reviews Neuroscience, 15,* 655–669.
Tamnes, C. K., Walhovd, K. B., Engvig, A., Grydeland, H., Krogsrud, S. K., Østby, Y., et al. (2014). Regional hippocampal volumes and development predict learning and memory. *Developmental Neuroscience, 36,* 161–174.
Townsend, E. L., Richmond, J. L., Vogel-Farley, V. K., & Thomas, K. (2010). Medial temporal lobe memory in childhood: Developmental transitions. *Developmental Science, 13,* 738–751.
Van Petten, C. (2004). Relationship between hippocampal volume and memory ability in healthy individuals across the lifespan: Review and meta-analysis. *Neuropsychologia, 42,* 1394–1413.
Watson, C., Andermann, F., Gloor, P., Jones-Gotman, M., Peters, T., Evans, A., et al. (1992). Anatomic basis of amygdaloid and hippocampal volume measurement by magnetic resonance imaging. *Neurology, 42,* 1743–1750.
Wechsler, D. (2011). *Wechsler Abbreviated Scale of Intelligence-Second Edition (WASI-II)*. San Antonio, TX: NCS Pearson.
Weiss, A. P., Dewitt, I., Goff, D., Ditman, T., & Heckers, S. (2005). Anterior and posterior hippocampal volumes in schizophrenia. *Schizophrenia Research, 73,* 103–112.
Wendelken, C., & Bunge, S. A. (2010). Transitive inference: Distinct contributions of rostrolateral prefrontal cortex and the hippocampus. *Journal of Cognitive Neuroscience, 22,* 837–847.
West, M. J., & Gundersen, H. J. (1990). Unbiased stereological estimation of the number of neurons in the human hippocampus. *Journal of Comparative Neurology, 296,* 1–22.
Wisse, L. E. M., Gerritsen, L., Zwanenburg, J. J. M., Kuijf, H. J., Luijten, P. R., Biessels, G. J., et al. (2012). Subfields of the hippocampal formation at 7T MRI: In vivo volumetric assessment. *Neuroimage, 61,* 1043–1049.
Yurgelun-Todd, D. A., Killgore, W. D. S., & Cintron, C. B. (2003). Cognitive correlates of medial temporal lobe development across adolescence: A magnetic resonance imaging study. *Perceptual and Motor Skills, 96,* 3–17.
Yushkevich, P. A., Wang, H., Pluta, J., Das, S. R., Craige, C., Avants, B. B., et al. (2010). Nearly automatic segmentation of hippocampal subfields in in vivo focal T2-weighted MRI. *Neuroimage, 53,* 1208–1224.
Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron, 75,* 168–179.
Zeithamova, D., & Preston, A. R. (2010). Flexible memories: Differential roles for medial temporal lobe and prefrontal cortex in cross-episode binding. *Journal of Neuroscience, 30,* 14676–14684.