frontiers in
AGING NEUROSCIENCE

# What-Where-When memory, unlike other cognitive abilities, is unimpaired in healthy people over 70.

Adele Mazurek, Rajameenakshi Bhoopathy, Jennifer C.A. Read, Peter Gallagher and Tom V Smulders



| | |
|---|---|
| Journal Name: | Frontiers in Aging Neuroscience |
| ISSN: | 1663-4365 |
| Article type: | Original Research Article |
| First received on: | 17 Oct 2014 |
| Frontiers website link: | www.frontiersin.org |

# What-Where-When memory, unlike other cognitive abilities, is unimpaired in healthy people over 70.

Adèle Mazurek, Rajameenakshi Bhoopathy, Jennifer C. A. Read, Peter Gallagher, Tom V. Smulders*

Institute of Neuroscience, Newcastle University, Newcastle upon Tyne, UK

Correspondence:

Dr. Tom Smulders
Newcastle University
Institute of Neuroscience
Henry Wellcome Building
Framlington Place
Newcastle upon Tyne
NE2 4HH
United Kingdom
+44-191-208-5790
tom.smulders@ncl.ac.uk

Running Title: Unimpaired WWW memory in aging

**Abstract**

Many cognitive abilities decline with aging, making it difficult to detect pathological changes against a background of natural changes in cognition. Most of the tests to assess cognitive decline are artificial tasks that have little resemblance to the problems faced by people in everyday life. This means both that people may have little practice doing such tasks (potentially contributing to the decline in performance) and that the tasks may not be good predictors of real-world cognitive problems.

In this study, we test the performance of young people (18-25 years) to that of 60-69-year-olds and 70+-year-olds on a novel, more ecologically valid test of episodic memory: the What-Where-When (WWW) memory test. We also compare them on a battery of other cognitive tests, including working memory, psychomotor speed, executive function, and episodic memory. Whereas both groups of older people show the expected age-related declines on most of the tests, only the 60-69-year-olds were impaired on the WWW memory test. They were less able to bind information about which object had been hidden to the location and the time at which it had been hidden than either the 18-25-year-olds or the 70+-year-olds.

One possible explanation of our results is that the 70+-year-olds may have represented a self-selected healthy cohort, while the 60-69-year-olds may have been more randomly selected and hence may have contained people with early signs of cognitive pathologies. If this is the case, then our more ecologically-valid memory test might be more sensitive to signs of early pathology, while most of the other cognitive tests we administered are susceptible to the effects of normal aging. This hypothesis remains to be tested. Because self-reported memory complaints also predicted performance on the WWW task, we suggest that the ecologically-valid WWW memory task may also be useful as a predictor of everyday memory abilities in people's natural environment.

**Key words:** episodic memory, aging, old-old, young-old, what-where-when memory, ecological validity, neuropsychology, dementia,

## 1. Introduction

Dementia is a degeneration of the brain and therefore of many cognitive processes, including memory. Memory deficits are often evident before any other signs of dementia are obvious (Masur et al., 1994;Bäckman et al., 2001;Jorm et al., 2005). Monitoring memory function can therefore be useful for early diagnosis of dementia, which in turn can help with the management of the disorder, potentially therapeutically slowing down the progression. For example, it has been shown that early deficits in episodic memory abilities can be indicative of the likelihood of a person developing Alzheimer's dementia later on in life (Bäckman et al., 2001). Episodic memory is our memory for personally experienced episodes from our own past, which we typically experience as "Mental Time Travel": a mentally re-experiencing of the episode in question (Suddendorf and Corballis, 1997).

One of the problems with using cognitive indicators as potential early-warning signals for dementia is that many cognitive capacities diminish as we get older. Aspects of verbal short-term memory (e.g. digit span) and vocabulary may decline rapidly in later-life, although processing speed, working memory and long-term memory are all known to decline steadily as we age (Hedden and Gabrieli, 2004). With regard to long-term memory, while semantic processes are relatively unaffected, episodic memory exhibits a much greater degree of decline (Nyberg et al., 2012). Numerous studies have shown performance impairments in episodic-like memory tests in older people, even if there is no evidence of dementia or Mild Cognitive Impairment (MCI; Harris et al., 2002). For example, Kessels et al. (2007) demonstrated broad performance decrements in older adults on a visuo-spatial episodic memory task which were especially pronounced in conditions requiring contextual binding. Tasks requiring the learning and recall of word lists (e.g. Rey-Auditory or California Verbal Learning Tests (R-AVLT/CVLT)) have been found to be impaired in aging (Lundervold et al., 2014), with particular deficits in temporal order indices (Blachstein et al., 2012). There is also some suggestion that the age-related decline in verbal episodic memory may be greater in males that females (Lundervold et al., 2014). Because of these changes, it is sometimes difficult to distinguish the early signs of dementia from natural declines in cognitive capacity with old age. However, it has been suggested that measures such as the Rey-AVLT may be useful in delineating different dementias (Tierney et al., 1994;Ricci et al., 2012).

One potential criticism of many of the clinical tests of episodic memory is that they do not have very high ecological validity (Sbordone and Long, 1996). Everyday episodic memory typically has a number of characteristics that are not easily captured in most clinical tests: it is made up of long-term memories for unique events in their spatiotemporal context (what happened, where it was, when it was). The information is usually encoded in an incidental manner, and freely recalled, without any cues relating to the original event. Laboratory tests usually match some of these features, but rarely all of them. For example, some tests, like the R-AVLT, are about free recall of long-term (30-min) memory (in this case of a list of words), but the information is just a list of words (no spatiotemporal context needs to be remembered, although the optional temporal-order trial can be administered; Vakil and Blachstein, 1994) and it is learned in an intentional manner and rehearsed several times. Other tests (e.g. the object relocation task; Kessels et al., 1999) capture the binding between objects (what happened) and spatial locations (where it was); they typically do this over short retention intervals, using recognition processes for the items (though not for the locations) and again include intentional encoding of the information. The advantage of all these tests is that the experimenter/clinician knows exactly which answers are correct and which are wrong, because they control the information to be retained. When more ecologically valid measures of episodic memory are used, such as having people freely recall real events from their own lives, the scoring of these memories necessarily has to rely on the amount of detail recalled, rather than on the accuracy of these memories, as no objective record usually exists of the original event (e.g. Irish et al., 2011).

Recently, a number of new tests have been developed to try and overcome some of the drawbacks of the traditional tests and gain more ecological validity. Some of these tests are based on a reconceptualization of episodic memory which was originally adapted for use with non-human animals. In the absence of language, the tests are based on the animal experiencing two unique episodes, and then demonstrating through their behavior what is remembered about these two episodes. These tests emphasize the long-term retention of unique information about events in their spatio-temporal context. In the first study to do so, food-hoarding California scrub jays (*Aphelocoma californica*) hid two types of food on each of two separate occasions. Having been trained to know that the preferred food type degrades after a several days, but the non-preferred one does not, they were then tested shortly after the second hiding episode. They only recovered the preferred food in the locations where they had hidden it in the second hiding episode, showing that they remembered which food (what) they had hidden in which locations (where) and on which occasion (when) (Clayton and Dickinson, 1998). Since then, several variations on this task have been developed for other animals, including other birds (Feeney et al., 2009;Zinkivskay et al., 2009;Gould et al., 2012), as well as rats (Eacott et al., 2005;Babb and Crystal, 2006;Roberts et al., 2008).

More recently, adaptations of these tasks have been developed for humans. In a typical task, participants experience one or two unique events, and then have to recall what happened where, and when (Plancher et al., 2010;Hayne and Imuta, 2011;Holland and Smulders, 2011;Russell et al., 2011;Russell and Hanna, 2012;Cheke and Clayton, 2013;Inostroza et al., 2013;Newcombe et al., 2014). This is either in terms of "in which of the two episodes", or "when in the episode", asking about the sequence in which things happened. In the current study, we use a further adapted version of the task first reported by Holland and Smulders (2011). In this task, participants hide 8 different objects in 8 different (pre-determined) locations in a real-world room on each of 2 occasions, separated by several hours. After another two hours, they are then taken back into the room and asked to recall which objects they had hidden where, and on which occasion. The participants are told a cover story about the study, so that they would encode the information incidentally, rather than intentionally. Therefore, this task tests long-term memory for incidentally-encoded information about unique episodes in their spatiotemporal contexts. Part of the memory retrieval is based on free recall, although the spatial locations are in view of the participant and can therefore be recognized, rather than recalled. Because the participants move around a real environment and interact with real objects and locations, the task has added ecological validity over computer-based or paper-based tests. Because the objects are all unique, the task also allows us to test object memory and spatial memory independently of the memory for how different features of the episodes are bound together. The goal of the study was to investigate whether older participants would show a deficit in this putative test of episodic memory, and compare their performance to other cognitive tasks in which age differences are well established.

## 2. Materials and methods

### 2.1. Participants

Fifty eight people participated in the study, which was approved by Newcastle University's Faculty of Medical Sciences Ethics Committee (approval number 00414), and run between January and May 2012. The sample was composed of three age groups: twenty-six young adults (17 women and 9 men, mean age 20, ranging from 18-24, all students), eighteen 60-69-year olds (10 women and 8 men, mean age 65, ranging from 61-69) and fourteen people over 70 (9 women and 5 men, mean age 77, ranging from 70-85). The split between "younger old" and "older old" participants is a common one in the literature, and there is no consistency as of the cut-off (e.g. Aslan and Bäuml, 2013;Mammarella et al., 2013;Yi and

Friedman, 2013;Zavagnin et al., 2014). One of the 70+-year-old participants had a visual impairment which prevented them from reading, so tasks that involved reading words or numbers were not administered to this participant. Each participant spoke English as a native language or spoke it fluently enough to study at a UK higher education establishment. All participants underwent the same procedure. At the end of the experiment, older people received a £20 gift card for a shopping center, students of the School of Psychology were given participation credit for their degree and other students were paid £5.

### 2.2. Procedure

Participants attended the lab twice in the same day. In the morning session, they were briefed on the procedures and filled out consent forms. They then performed the first session of the What-Where-When task, adapted from Holland and Smulders (2011). They then went away for approximately 2 hours, during which they had lunch. After lunch, they first performed the second session of the What-Where-When task. They were then run through a battery of other neuropsychological tasks, before being tested for their memory in the final What-Where-When session. Details about the exact procedures for the different tasks can be found below.

#### 2.2.1. What-Where-When task

The WWW task was conducted with all participants unaware that they were participating in a memory task. They were told that the aim of the study was to investigate how well they could repeat a sentence ("She bought a bit of butter") again and again under distracting conditions, and whether practice improved their performance. They were made to believe that their voice was being recorded. In addition to being part of the cover story, the sentence also served as articulatory suppression (Hanley, 1997), to prevent participants from verbally rehearsing any information during the task. In the first session, participants were required to hide 8 objects (*an earring, a spoon, a coin, a pencil top, a toy frog, a party blower, a fold-back paperclip and a playing card*) in pre-determined locations in a cluttered office room. The objects were given to the participant one at a time, and the locations were identified during the task by the experimenter pointing at the locations for the participant to place an object in.

The second session occurred in the afternoon, on average two hours after the first session. First, participants were required to perform the same procedure as in the morning session with eight other objects (*a key, a plastic ball, a clothes peg, a rubber band, a bottle cap, an eraser, a top and a toy snake*) in eight new pre-determined locations. Finally, after having been tested on all the other neuropsychological tests (see below), the participants were returned to the room in which they had hidden the objects, and asked to recall which objects they had hidden in which locations and on which of the two occasions. They were also encouraged to report any incomplete information they could recall (e.g. items for which they could not remember the locations or vice versa). After they had recalled all the information they could, they were asked how they experienced the recall of the information: whether they re-experienced the hiding events in their heads ("remember"), or whether they just knew the information ("know"). They were also asked how vividly they re-experienced the information on a scale from 1-5, based on the Vividness of Visual Imagery Questionnaire (Marks, 1973).

**2.2.2. Memory self-assessment**

Right after the second hiding session, participants filled in three self-evaluation questionnaires: the Memory Complaint Questionnaire (MAC-Q; Crook et al., 1992) and Every Day Memory questionnaire (Sunderland et al., 1983) were used to assess perceived memory problems and the Geriatric-Depression-Scale questionnaire (GDS) was used to assess the general mood of the participants (Greenberg and Kurlowicz, 2007).

Then, a battery of neuropsychological tests was performed. The exact order was designed such that shorter tests could be run during retention intervals of the longer tests. We present the tests here in order of their complexity.

**2.2.3. Rey Auditory Verbal Learning Test (R-AVLT)** (Rey, 1964).

Participants listened to a list of 15 words (one word every 2 seconds; List A), which had been recorded using Audacity 1.3 beta by a native English speaker. They were then asked to immediately recall this list (measure A1). After this, a learning phase was carried out during which participants were presented with the list four more times and after each presentation they were again asked to verbally recall the list (A2-A5). Immediately after the fifth recall, participants were required to memorize a new list of 15 words (List B) and asked to immediately verbally recall them (B), followed by another verbal word recall from List A (A6). The output of this test was a measure of retroactive interference (RI = A5-A6) and proactive interference (PI = A1-B) scores. Then, around 30 minutes later participants were again required to recall the words from list A (A7).

Following the delayed word recall there was a word recognition task of the 30 words from List A and List B. The participants were presented with 50 words (the 30 words from lists A and B, plus 20 new words), and taken through this list by the experimenter. For each word, they needed to identify whether it was a new word or not, and if not, which list it belonged in. Temporal order judgment assessment followed the recognition trial: participants had to reorganize 15 pieces of paper on which the words of list A had been written in the correct order. The same procedure was used for the words of list B. We used three different measures of how well the reconstructed order matched the original order: (1) Hits: the number of word correctly place at their original serial position (2) Absolute deviation: this score was calculated by summing the absolute deviation of each word from its original position. The score for each scoring deviation ranges from 0 to 14 (3) Correlation: Pearson product-moment correlation calculated for each subject, between the listed order and the true order (Vakil and Blachstein, 1994).

**2.2.4. Object Relocation** (Kessels et al., 1999).

This paradigm is made up of 5 different test conditions: an Object Recognition Memory (ORM), in which participants have to memorize and then pick out 10 objects (from a choice of 20); a Visual Spatial Reconstruction (VSR), in which a spatial array of identical objects is shown on one side of a computer screen, and the participants have to copy it on the other side of the screen; a Position Only Memory (POM), in which 10 identical objects are presented on the screen for the participant to memorize, and then reconstruct after a retention interval; an Object Location Binding (OLB), in which 10 different objects are presented on the screen to be memorized, which then need to be matched to indicated locations after a retention interval;

and the Combined Object Memory (COM), which is a combination of POM and OLB, in that 10 objects and locations need to be memorized, and the locations are not shown after the retention interval. For every condition, there was first a practice trial with fewer objects/locations, followed by two full trials with 10 items each. For the memory versions of the task (ORM, POM, OLB and COM), we had one trial with a zero-second retention interval, and one with a three-minute retention interval. Half the participants did the short retention interval first, and half did the long retention interval first. The outcome measures for the ORM and OLB are the number of correctly identified objects/locations, whereas for the other three tasks, the outcome measure is the sum of the absolute distances between the objects and their correct locations (or in the case of the POM, the nearest correct location).

#### 2.2.5. Standard neuropsychological tests.

Verbal working memory was tested using the Forward Digit Span, while verbal working memory combined with executive function was tested using the Backward Digit Span (Wechsler, 1981;Lezak et al., 2004). We used the maximum span remembered as the outcome measure for both tests. Visual working memory was tested using the CANTAB (Cambridge Cognition, Cambridge, United Kingdom) version of the Corsi Block task (Spatial Span – SS), the CANTAB Paired Associates test and the Visual Patterns Test (Della Sala et al., 1997). Psychomotor speed was tested using the Trail making Test A, and psychomotor speed plus executive function using the Trail making Test B (Lezak et al., 2004). Finally, language comprehension was tested using two subtests from the Speed and Capacity of Language Processing (SCOLP) test: the SCOLP Word and the SCOLP Comprehension tests (Baddeley et al., 1992).

## 2.3. Data analysis

### 2.3.1. Classic statistics

All data analyses (except for the Bayes Factor calculations, see below) were performed in IBM® SPSS® v21. For normally distributed interval data, we used a General Linear Model (GLM) approach, which gives classic F-values as the output. For counts of correct responses (e.g. SCOLP, AVLT, WWW), we used the Generalized Linear Model (GzLM) approach, with data from a binomial distribution with logit link function; for repeated measures of the same, we use the Generalized Estimating Equations (GEE), with the same link function, and an unstructured correlation matrix. The GzLM and GEE give Wald's $\chi^2$ as the output statistic. All models were simplified by removing non-significant interactions, starting with the highest-level interactions. For CANTAB errors, we used GzLM with data from a Poisson distribution with log-link function and for the Vividness scale; the data was treated as ordinal, using a log-link function. Results were considered significant at an $\alpha$-level of 0.05.

### 2.3.2. Bayes Factor

When differences between groups are not significant, this can be because of a real absence of a difference, or because of a lack of statistical power to detect a difference. One way to distinguish between these two options is to calculate a Bayes Factor (Dienes, 2011), which calculates how much more likely a given hypothesis is to be correct, given the data obtained.

A Bayes Factor above 1 indicates that confidence in the hypothesis should increase, whereas a Bayes Factor below 1 suggests it should decrease. Online calculators exist to calculate Bayes Factors for comparisons of continuous variables between two groups (Dienes, 2011). However, the main dataset to which we wanted to apply the calculation was the outcomes of the What-Where-When task, which consists of binary data (correct or not for each item, location or combination). We therefore designed our own Bayes Factor calculator for binary data.

To calculate Bayes Factors, our model assumes that the success probability (probability of getting a trial correct, for a given definition of correct in that particular analysis) is affected only by the grouping of interest, and is the same for all trials conducted by all subjects in a given group. This allows us to use simple binomial statistics. In the text below, we will speak about "young" and "older" groups, but the principle applies to any two groups being compared to each other.

We write $N_Y$ for the total number of trials completed by all Young participants, and $M_Y$ for the number of these which were successful; $N_O$, $M_O$ are analogous quantities for Older participants. The observed difference in the proportion of successful trials between these two age-groups is then

$$D = M_Y/N_Y - M_O/N_O,$$

where a positive difference means that young people did better. We assume the mean success probability, averaged over both age-groups, is the observed proportion of successes when both groups are combined:

$$\mu = (M_Y+M_O)/(N_Y+N_O).$$

By definition, the underlying (population) success probability for younger participants is then $\pi_Y = \mu+\Delta/2$, and that for older participants is $\pi_O = \mu-\Delta/2$. Clearly, both of these probabilities must lie in the range [0,1]. Thus, the assumed mean success probability, $\mu$, constrains the possible values of the true difference $\Delta$. For example, consider the most extreme situation when the true probability is 0 for one group. In order to get a mean probability of $\mu$ for both groups, the true probability for the other group must be $2\mu$. It cannot go above this and still keep the mean probability at $\mu$, since the probability for the first group cannot be negative. These considerations imply that $\Delta$ must lie between $\pm\Delta_{lim}$, where $\Delta_{lim}=2\mu$ for $0<\mu\leq0.5$ and $\Delta_{lim}=2(1-\mu)$ for $0.5<\mu\leq1$.

We now want to compute the likelihood of the observed (sample) difference in the proportion of successful trials, D, if the underlying (population) difference in success probability is really $\Delta$. We write this as $\Pr(D|\Delta)$. To calculate this, we consider all the possible scores which would give the observed value of D, given the number of trials actually performed by each age-group. The probability of each score is given by the simple binomial distribution:

$$\Pr(m|N,\pi) = \frac{N!}{m!(N-m)!}\pi^m(1-\pi)^{N-m}$$

We sum the product $\Pr(m_Y|N_Y,\pi_Y)\Pr(m_O|N_O,\pi_O)$ over all pairs of $(m_Y,m_O)$ which satisfy $0\leq m_Y\leq N_Y$, $0\leq m_O\leq N_O$ and $(m_Y/N_Y - m_O/N_O) = D$. This is our estimate of $\Pr(D|\Delta)$ : the probability of observing a particular difference D in the proportion of successful trials, given an actual difference Δ in the probability of a successful trial.

To compare the null hypothesis that there is no difference in success probability between age-groups, Δ=0, with the experimental hypothesis that Δ could be non-zero, we computed the Bayes Factor, B. This is the ratio of the likelihood of the observed difference D under the experimental hypothesis to its likelihood under the null hypothesis, $B=L_{expt}/L_{null}$. These are

$$L_{null} = \Pr(D|\Delta = 0)\ ; \qquad L_{expt} = \int_{-\Delta_{lim}}^{+\Delta_{lim}} d\Delta\, P(\Delta)\Pr(D|\Delta)\ ,$$

where Pr(D|Δ) is calculated as described above and P(Δ) is *the a priori* distribution for Δ under the experimental hypothesis. In our analysis, we set this distribution as a half-gaussian in the direction of younger people doing better:

$$P(\Delta) = \frac{1}{2\sigma\sqrt{2\pi}}\exp\left(-\frac{\Delta^2}{2\sigma^2}\right) \text{ for } \Delta>0 \text{ and } P(\Delta)=0 \text{ otherwise.}$$

The standard deviation was set to half of the maximum possible difference in success probability between the two groups, $\sigma=\Delta_{lim}/2$. This sets the maximum difference at the 95% confidence interval, as suggested by Dienes (2011). From the expressions above, this means that σ=μ or (1-μ), whichever is smaller. In our data-set, μ<0.5, so the standard deviation σ of our prior distribution for the difference in probability between the groups is equal to the estimated mean probability across both groups.

MATLAB code for this analysis is available at http://www.jennyreadresearch.com/research/matlab-code/bayes-factors-for-binomial-data/

# 3. Results

## 3.1. Memory self-assessment

Both groups of older people reported fewer memory problems on the Everyday Memory Questionnaire (EMQ) than younger people ($F_{2,54}$=16.96, P<0.001; Fig. 1A), while people who reported more everyday memory problems also reported a lower mood on the Geriatric Depression Scale (GDS) (covariate in the GLM model; $F_{1,54}$=6.09, P=0.017). There were no age differences in scores on the GDS ($F_{2,55}$=0.042, p=0.959) and the effect of mood on EMQ did not differ between the three age groups, so the non-significant interaction between age and GDS was left out of the GLM model. In contrast, in the Mac-Q, elderly people describe their memory as being poorer now than high school or college, more so than young people ($F_{2,54}$=9.52, P<0.001; Fig. 1B), and there was no effect of current mood on this self-report of memory performance (covariate; $F_{1,54}$=0.02, P=0.888). Again, the non-significant interaction between age and GDS was left out of the model. According to the criteria of Crook et al. (1992), a Mac-Q score ≥ 25 is associated with memory decline. By this standard, more than half (17/32) of elderly people were affected by age associated memory decline, while none of the young people were so affected ($\chi^2_2$=19.73, P<0.001). The proportions of 60-69-year-old (9/18) and 70+-year-old (8/14) people so affected were similar ($\chi^2_1$=0.16, P=0.688).

### 3.2. Working memory, executive function and knowledge.

Participants were tested using a battery of measures for which age differences were expected based on previous literature. This served to verify that the sample was similar to previous samples of younger and older people. Older people performed worse on the visuospatial working memory tests (Fig. 1C): the Spatial Span test ($F_{2,55}$=44.96, P<0.001), the Visual Patterns Test ($F_{2,55}$=16.76, P<0.001) and the CANTAB Paired Associates test ($\chi^2_2$=205.077, P<0.001). For both the Spatial Span and the Visual Patterns test, 60-69-year-old and 70+-year-old participants did not differ in their performance, but for the CANTAB Paired associates, the 70+-year-olds performed even worse than the 60-69-year-olds ($\chi^2_1$=15.33, P<0.001). Whereas the three groups do not show a significant difference in performance on the Forward Digit Span (a test of verbal working memory; $F_{2,55}$=0.82, P=0.445), older people perform worse than younger people on the Backward Digit Span, a test of executive function ($F_{1,56}$=5.38, P=0.024). When the groups were separated into 60-69-year-olds and 70+-year-olds, the difference with young people was only significant for the 70+-year-olds (p=0.022). As expected, older people were slower on Trail Making A, a test of psychomotor speed ($F_{2,54}$=9.21, P<0.001; Fig. 1D), with no differences between 60-69-year-olds and 70+-year-olds (p=0.733). Controlling for psychomotor speed by using the time taken to complete Trail Making A as a covariate in the analysis of Trail Making B, a test of executive function, again indicates that older people perform worse on executive function than younger people ($F_{2,53}$=4.30, P=0.019), with no difference between the two groups of older people (P=0.785).

In contrast to measures of speed, working memory and executive function, older people outperformed younger people on the SCOLP tests of vocabulary ($\chi^2_2$=92.88, P<0.001) and sentence comprehension ($\chi^2_2$=9.34, P=0.009). There were no significant differences between the 60-69-year-old and the 70+-year-old group. There were no age differences in the time in which participants finished the sentence comprehension task ($F_{2,54}$=1.52, P=0.229).

### 3.3. Rey-AVLT

#### 3.3.1. Word recall and recognition

In order to compare the learning and forgetting curves for the two age groups, a GEE analysis was performed with the different stages of the R-AVLT as within-subject factor and age as between-subject factor (Fig. 2A). Older people remembered fewer words than younger people ($\chi^2_2$=43.47, P<0.001), but there was no difference between 60-69-year-old and 70+-year-old participants (P=0.879). As expected, the number of words recalled increased from A1 to A5, and decreased from A5 to A7 ($\chi^2_6$=427.73, P<0.001). The change over time was different for the age groups (interaction: $\chi^2_{12}$=39.03, P<0.001).

As they did not differ from each other, data from the 60-69-year-old and 70+-year-old participants were pooled, and the patterns over the different steps of the R-AVLT were examined in more detail. The learning curves from A1 to A5 were analyzed first. Older people consistently remembered fewer words than younger people ($\chi^2_1$=39.66, P<0.001), and both groups improved with repetition ($\chi^2_4$=381.27, P<0.001). Again, the interaction between age and learning was significant ($\chi^2_4$=31.55, P<0.001), indicating that the change in performance was different between the older and the younger participants. Indeed younger participants did not significantly improve anymore from A4 to A5 (post-hoc pairwise comparisons, $P=1.00$), whereas older participants continued to improve (Fig. 2A).

The effect of the Retroactive Interference (of having list B between A5 and A6) was then examined. The age difference remained overall ($\chi^2_1=28.61$, $P<0.001$), and there was a significant retroactive interference effect ($\chi^2_1=34.81$, $P<0.001$), but the interaction between the two factors did not quite reach significance ($\chi^2_1=2.98$, $P=0.084$). However, if the difference scores between A6 and A5 were examined using a General Linear Model, the Retroactive Interference effect is much stronger in the older group ($F_{1,56}=16.56$, $P<0.001$; Fig. 2A). Interestingly, although there clearly was an overall Proactive Interference effect of list A when retrieving list B ($F_{1,56}=6.27$, $P=0.015$), there was no significant difference between young and old people in this effect ($F_{1,56}=0.003$ $P=0.959$).

Finally, the forgetting from A6 to A7 was investigated. Whereas younger people continued to outperform older people ($\chi^2_1=27.67$, $P<0.001$), and forgetting indeed occurred ($\chi^2_1=11.21$, $P=0.001$), this forgetting did not differ between the two age groups ($\chi^2_1=2.70$, $P=0.10$). In this case, this lack of an age difference in forgetting was confirmed by the General Linear Model comparing the difference scores between A6 and A7 ($F_{1,56}=0.408$, $P=0.526$).

Participants were also asked to recognize the words from list A and list B in a larger list with 20 foils. d-prime was calculated for both lists, based on the number of hits (correctly recognized words) and false alarms (words attributed to the list that were not part of the list; Fig. 2B). Performance was much better for list A than for list B ($F_{1,55}=159.09$, $P<0.001$) for all age groups. Younger participants outperformed older participants ($F_{2,55}=6.81$, $P=0.002$), and but only for list A (interaction: $F_{2,55}=10.53$, $P<0.001$), although this may be due to a floor effect for performance on list B. There was no difference in performance between the two older groups ($P=0.493$).

### 3.3.2. Word order

The temporal order in which things happen is often cited as a crucial component of episodic memory. We had three measures of temporal order in recalling the word lists in the Rey-AVLT: Hit score (number of items in the correct position; Fig. 2C), absolute deviation from correct position for each item (Fig. 2D) and correlation between the real position and the recalled position (Fig. 2E). We conducted either a GEE (hits) or an RM ANOVA (absolute deviation and correlation coefficients) with scores on list A vs. list B as the within-subjects factor and age as the between-subjects factor. Older people performed worse than younger people (lower hit scores: $\chi^2_2=6.58$, $P=0.037$; higher absolute deviation: $F_{2,54}=6.23$ $P=0.004$; lower Pearson correlation: $F_{2,54}=5.47$ $P=0.007$), and in no cases did the two groups of older people differ from each other. For all ages, performance was better for list A than for list B (higher hit score: $\chi^2_1=118.40$, $P<0.001$; lower absolute deviation: $F_{1,54}=286.12$, $P<0.001$; higher Pearson correlation: $F_{1,54}=188.33$, $P<0.001$). There was a significant interaction between age and list for Hit score ($\chi^2_2=14.97$, $P=0.001$), but not for absolute deviation or correlation (Absolute deviation: $F_{2,54}=0.87$, $P=0.424$; correlation: $F_{2,54}=0.62$, $P=0.544$). For the hits, it is possible that the age difference only exists for list A, not for list B. However, we should be cautious with this interpretation, as this could be a floor effect for list B (<2 hits for all groups).

### 3.4. Object Location Task

In the Visual Spatial Reconstruction task, younger people perform better than older people ($F_{2,54}=$ 6.69, P=0.003; Fig. 3A), with no difference between the two groups of older people. Because of this age difference in visuo-spatial perception, performance on VSR was controlled for when investigating age differences in spatial memory (POM and COM), by using the average VSR score across the two sessions as a covariate in the analysis. Thus controlling for worse spatial perception, no age differences were found in either Place Only Memory ($F_{2,53}=$ 0.77, P=0.469; Fig. 3B) or Combined Object Memory ($F_{2,52}=$ 1.24, P=0.298; Fig. 3C). There was also no difference between the two delay conditions in either measure (POM: $F_{1,53}=$ 0.757, P=0.388; COM: $F_{1,52}=$ 0.08, P=0.774). Age differences were found in the Object Recognition Memory task ($\chi^2_2$=9.89, P=0.007; Fig. 3D) and the Object Location Binding task ($\chi^2_2$=10.58, P=0.005; Fig. 3E). In both cases, younger people outperform 60-69-year-olds, with 70+-year-old performance in between, and not significantly different from either other group. Again, delay did not significantly affect performance on either of these two measures (ORM: $\chi^2_1$=0.03, P=0.863; OLB: $\chi^2_1$=3.18, P=0.075).

### 3.5. Performance on the WWW task

#### 3.5.1. Overall performance

Performance on the integrated *What-Where-When* measure differed among the age groups. Interestingly, the 60-69-year-olds performed significantly worse than either the young or the 70+-year-olds ($\chi^2_2$=12.96, P=0.002): whereas the young and the 70+-year-olds remembered on average 2.12$\pm$0.27 and 2.50$\pm$0.39 WWW combinations respectively, the 60-69-year-olds remembered only 0.94$\pm$0.22 correct combinations (Fig. 4). Given these data, it is 8.3 times more likely that there is no difference between young and 70+-year-old participants than that younger people outperform the 70+-year-old group (for Bayes Factor calculation, see Methods and Supplementary material), suggesting this lack of difference is not due purely to a lack of statistical power (Jeffreys, 1961;Dienes, 2011).

Memory for incomplete combinations of *What-Where*, *What-When* and *Where-When* (not including the correct WWW combinations; Fig. 4) was then examined. Interestingly, there were no significant age group differences in the performance on these combinations ($\chi^2_2$=2.62, P=0.270). The performance on the different combinations was very different, however ($\chi^2_2$=140.89, P<0.001). Few participants recalled any incomplete *What-Where* combinations (n=50 did not recall any, n=7 recalled 1 and n=1 recalled 2), implying that when people recalled which object was hidden where, they also remembered on which occasion that had happened. Participants recalled more incomplete *What-When* combinations (on average 10%$\pm$1.2% of the combinations they had not recalled as a full WWW combination), and even more incomplete *Where-When* combination (on average 28%$\pm$1.9% of the combinations not recalled as full WWW combinations). This strongly suggests that it is possible and even likely to bind objects or locations to time frames by themselves, but when both object and location are recalled, the time frame is recalled as well. This pattern of performance across the three types of incomplete combinations did not differ significantly across age categories (interaction: $\chi^2_4$=2.37, P=0.668; Fig. 4).

Finally, performance on remembering individual objects or locations that had not been recalled as part of a combination of any kind was investigated. Similar to the incomplete combinations with *When*, Locations were remembered much more commonly than objects (20%±1.8% of the locations not recalled in combination vs. 8%±1.1% of the objects not recalled in combination; $\chi^2_1$=41.51, P<0.001). There were no differences among the age categories ($\chi^2_2$=0.52, P=0.773), nor was there an interaction with age ($\chi^2_2$=0.19, P=0.911; Fig. 4).

The lack of significant age differences in the incomplete combinations and individual items could be due to a genuine absence of age differences, or due to lack of statistical power. In order to determine whether there really is no age difference, Bayes Factors were calculated for each of these 5 comparisons between young people and both groups of old participants. In this study, the Bayes Factors for all these comparisons indicated that it was 3.8 to 19.6 times more likely that there really are no age differences than that the younger people perform better than the older people, suggesting the lack of difference is not due to a lack of statistical power. One exception is the comparison of the incomplete What-Where combinations, where no conclusion could be drawn due to the small number of responses in that category. In conclusion, while there is a 'genuine' difference between young and 60-69-year-old participants in remembering full WWW combinations, this is not the case for the incomplete combinations.

### 3.5.2. Subjective experience of WWW recall

In all age groups, participants claimed to 'relive the session in their head' ("remember") significantly more often than to just know ("know") which objects were hidden where and when ($\chi^2_1$=10.38, P=0.001; in total n=42/58), and this did not differ among the age groups ($\chi^2_2$=2.27, P=0.321). People who claim to relive the sessions also scored their recall experience higher on the vividness scale ($\chi^2_1$=6.11, P=0.013). There were no overall age group differences on the vividness scale ($\chi^2_2$=4.72, P=0.094), but there was a much larger difference in vividness between "remember" and "know" in the 70+-year-old group than in the other two groups (interaction: $\chi^2_2$=6.22, P=0.045; Fig. 5A).

Whether the mode of recall affected accuracy in the recall of the full *WWW* combinations was then investigated (including age category and the interaction between mode of recall and age also in the model). There was no evidence that mode of recall significantly affected recall of the full *WWW* combinations ($\chi^2_1$=2.10, P=0.147; interaction with age: $\chi^2_2$=4.49, P=0.106; Fig. 5B). Increasing vividness of experience did not significantly improve memory outcomes ($\chi^2_1$=1.82, P=0.177).

Another way to approach the mode of recall is to investigate the order in which the information is recalled. A retrieval order that follows the order of the original experience might indicate a mental time travel strategy. The correlation between the order of recall of hiding locations and the order of hiding in those locations was therefore examined. This correlation did not differ among the age categories ($F_{2,53}$=0.89, P=0.419), nor did it differ from zero across all participants (Intercept: $F_{1,53}$=0.26, P=0.609), suggesting people are not following their original route mentally when recalling the information. The average number of ranks (absolute difference) that any given recalled location was from its original rank also did not differ among the age categories ($F_{2,54}$=0.54, P=0.585).

### 3.5.3. Rey-AVLT and WWW recall

Rey-AVLT and WWW are both purported measures of episodic memory. If this is the case, then individual variation in the each of the tasks should correlate across individuals. In order to test whether performance on the *WWW* combination was predicted by memory performance on the Rey-AVLT, another GzLM analysis was performed. Performance on a long-term memory task is dependent both on how much information was encoded in the first place, and how well this information is retained. For that reason, three measures from the Rey-AVLT were used to predict performance in the WWW task: the first was the total number of words recalled after a single exposure (A1), because in the WWW task, there was only one exposure to the information. The second was the number of words forgotten from A5 to A6 (A5-A6; Retroactive Interference, as the B list was learned between these two), and the final one was the number of words forgotten across the 30-min retention interval from A6 to A7 (A6-A7). The GzLM used these three variables as covariates and Age category as a fixed factor. Non-significant interactions between age and the three covariates were removed from the analysis in a stepwise manner until none remained. As reported above, the 60-69-year-olds performed worse than the young and the 70+-year-olds ($\chi^2_2$=12.08, P=0.002). People who could memorize more words in one exposure also remembered more WWW combinations ($\chi^2_1$=6.98, P=0.008; Fig. 6A), as did people who forgot fewer words from A6 to A7 ($\chi^2_1$=7.98, P=0.005; Fig. 6B). There was no significant effect of retroactive interference on remembering WWW combinations (A5-A6: $\chi^2_1$=1.86, P=0.173; Fig. 6C).

### 3.5.4. Object Location Memory and WWW recall

Object Location Memory is another purported episodic memory task that should measure similar processes to the WWW task, and hence predict performance on the WWW task. Because there were no effects of memory delay on any of the outcome measures from the Object Location Memory task, mean performance across the two trials of each type for each participant was calculated. For the POM and COM measures, performance on the VSR was controlled for by calculating the residuals from a regression against VSR, and then adding mean performance across all participants to those residuals, in effect calculating the memory performance while keeping VSR performance constant. Using these 4 measures as covariates, only COM significantly predicted WWW memory performance ($\chi^2_1$=7.25, P=0.007), with individuals with more accurate object relocation performance being better in the WWW memory task (Fig. 6D). This effect did not cancel out the age difference on the WWW memory task ($\chi^2_2$=9.54, P=0.009), indicating that both effects are independent of each other.

### 3.5.5. Self-reported memory problems and WWW recall

Finally, the question of whether self-reported memory problems in the Mac-Q and Everyday Memory Questionnaire (EMQ) predicted performance on the *WWW* task was explored. Using a similar analysis as above, people with a higher Mac-Q score (i.e. higher perceived memory problems) recalled fewer complete *WWW* combinations ($\chi^2_1$=4.03, P=0.045; Fig. 6E), and this did not interact with age category ($\chi^2_2$=0.53, P=0.768). The scores on the Every Day Memory Questionnaire did not predict performance on the WWW test ($\chi^2_1$=1.70, P=0.192), nor were there any significant interactions (Fig. 6F).

## 4. Discussion

There are two main findings from this study. Firstly, the WWW memory task is a valid measure of episodic memory, as performance on the task is predicted by two other episodic memory tasks (RAVLT and Object Location Memory), independent of the age effects on the tasks. Secondly, whereas the healthy 70+ group performed similarly to the 60-69 group (and worse than the younger participants) on most tasks, they performed much better (and not different from the younger group) on remembering complete What-Where-When combinations.

### 4.1. The WWW memory task measures episodic memory performance

Most participants reported using a mental time travel strategy ("remember"), rather than a semantic strategy ("know") to recall the information in the WWW memory task. Using this mental time travel strategy significantly improved performance of the 70+-year-olds over not using it. Additionally, performance on the WWW combination memory task was predicted both by how many words participants could learn in one exposure to the word list (AVLT A1) and by how well they can retain the list over a 30-min retention interval. This suggests that the WWW memory integrates initial learning with long-term retention of information, key features of episodic memory. Performance on WWW memory was also predicted by the COM error score (controlled for visuospatial perception). This is not completely surprising, as the two tasks have very similar requirements: remembering the binding of objects to locations, and having a view of the potential locations at the time the memory recall is tested. Finally, the level of self-assessed age-dependent memory problems (MAC-Q) predicts performance on the WWW memory task. Interestingly, this was not the case for the Everyday Memory Questionnaire. However, this instrument's value should be questioned in our study, because younger people reported more problems on this questionnaire than did older people (maybe because older people did not recall as many memory problems). These findings therefore suggest that the WWW memory task draws on similar processes to other episodic memory tasks. The design of the task (remembering real objects, incidentally memorized in a real-world environment) additionally increases its ecological validity, as is indicated by the prediction by the MAC-Q.

### 4.2. WWW binding is diminished in 60-69-year-olds, but preserved in 70+-year-olds

Like in many other studies, we found that older people performed worse than younger people in a battery of cognitive tests, including visual and verbal working memory, executive function, psychomotor speed, and a classic episodic memory test (RAVLT). In contrast, they performed better on semantic knowledge tasks, such as vocabulary. All this is similar to what we already know about cognitive aging (Hedden and Gabrieli, 2004). In the vast majority of these tests, there was no evidence of any differences between 60-69-year-olds and 70+-year-olds. This rules out the possibility that we inadvertently recruited 70+-year-olds of higher cognitive ability than the 60-69-year-olds.

However, in remembering What-Where-When combinations, the 70+-year-olds performed as well (if not better) than the younger people, while the 60-69-year-olds were significantly impaired. This lack of difference between 70+-year-olds and younger participants was not due to a lack of statistical power (as evidenced by the Bayes Factor and the fact that their average performance was in fact higher than the younger people), nor was it due to a lack of sensitivity of the outcome measure (as the difference between younger people and 60-69-year-olds was obvious: about 1 out of 16 combinations for the 60-69-year-olds, and 2-3 for the other two groups). So why were the 70+-year-olds spared in our study, while the 60-69-year-olds were not? One possible explanation is a difference in the sampling of the two populations. All older people were recruited through a database of older volunteers held at the Institute of Aging and Health, as well as through a more general database of volunteers held at the Institute of Neuroscience at Newcastle University. Many people in their 60s would not have retired yet and therefore be recruited from employees of the university or recent retirees. This sample is therefore likely to be relatively random with regards to general health and mental health, and may include people with very early signs of aging-related neural pathology (although we did not formally test for this; Hedden and Gabrieli, 2004). On the other hand, all 70+-year-old participants would have been retired for a while and therefore had to make a special effort to come and participate in our study. It is very likely that only the sub-group of 70+-year-olds who feels mentally and physically fit enough for the challenge would have volunteered for a study that was advertised as taking more than 4 hours. It is therefore possible that by recruiting volunteers, we have inadvertently only recruited people over 70 who are mentally healthy and feel up to the challenge of a cognitive test battery. Plancher et al. (2012) found that healthy 70+-year-olds outperformed patients with amnestic Mild Cognitive Impairment and with Alzheimer's Disease on a virtual WWW memory task. This adds plausibility to our hypothesis.

That being said, these same 70+-year-olds who did not show any deficits on the WWW task still showed all the same deficits on almost all the other cognitive tasks we set them, including a test of episodic memory (RAVLT). This indicates that they were not an exceptional sample of older people compared to any other studies published to date. So what is special about the WWW binding in the WWW memory task that they are spared from this decline? One thing that distinguishes our WWW test from all others is that it has much higher ecological validity. Whereas most neuropsychological tests require people to sit down with pen and paper or in front of a computer and effortfully memorize information or complete a task under time pressure, the WWW task is a real-world task, in which information was encoded incidentally, rather than intentionally.  The ecological validity might make it easier for older people to apply more efficient or effective strategies that they have honed in everyday life (Hedden and Gabrieli, 2004). This would be less likely to be the case for more typical neuropsychological tests, including the RAVLT.

Interestingly, there were no age effects on remembering individual objects or locations, nor of combinations of object with episode or location with episode. Only when an object had to be linked to a location (which in almost all case was also linked to the correct episode; see also Russell and Hanna, 2012), was there an age deficit. Interestingly, the only other task where 60-69-year-olds did worse than the younger people, but the 70+-year-olds did not (although it was also not significantly different from the 60-69-year-olds in this task), was the Object Location Binding task and the Object Recognition Memory task. OLB is also quite similar to the WWW task, in that the participants can see the locations, but need to bind the objects to them. This suggests that the main feature that is sensitive to early-onset aging-related pathology, but potentially not to healthy aging, is binding objects to locations. Our WWW findings are similar to those of Plancher et al. (2010), who tested people's memories of experiences in a virtual environment. They found that in their incidental encoding condition (but not in the intentional encoding condition), older people did not differ in object memory, but did have a deficit in the amount of binding between different elements they remembered. They did find age differences (even in the incidental encoding condition) in spatial memory performance (which we did not). However, their spatial memory performance was assessed differently from ours. Whereas our participants were put back in the same environment, and could use spatial cues to trigger their memories, in their case, the participants were asked to describe where different features occurred along a virtual road through a virtual town, and to draw a map of this virtual road. Their memory testing was therefore completely free recall, whereas ours was not. We will test the difference between free recall and cued recall in a future study to ascertain the effect of the assessment method on performance and on the age-sensitivity of this performance. Whereas their virtual environment allows for potentially more variation in the to-be-remembered information, they had to exclude 15 older people because they did not feel comfortable in the virtual driving task. Our real-world task avoids such complications and is simpler to administer in any setting.

Our findings suggest the hypothesis that object-location binding might be a sensitive way to distinguish healthy aging from early age-related pathologies. Doing this in an ecologically-valid manner like in our WWW memory task may additionally allow older people to utilize their real-world skills for dealing with their environment, which would make the outcomes more predictive of their real-world capacities, which is a crucial aspect of assessing memory in older people. Unlike a similar approach by Plancher et al. (2010, 2012), our real-world approach does not require specialized software or requiring people to become comfortable with navigation in a virtual environment. We do believe that the WWW test may be made more sensitive by increasing performance of the younger group to give more dynamic range to the outcome measures. In this particular version of the task, the participants did not remember many of the objects, whether in combination with location or not (fewer than half of the objects were ever recalled, even by young participants). This is probably because participants were given the objects in their hands and told to hide them in the indicated locations. Because they believed the objects to be a distractor, they may not have paid much attention to what was put into their hands. Object memory may be improved by making people select the objects themselves, guided by a list of pictures next to the pile of objects. This might then also increase the total number of correct WWW combinations recalled.

### 4.3. Conclusion

Memory for the binding of objects with locations (and occasions) in a long-term incidentally-encoding memory task was sensitive to aging in a relatively randomly selected population of older people, but was not affected in a self-selected mentally healthy population of 70+-year-olds. This opens up the possibility that the WWW memory task could be insensitive to the normal cognitive declines of aging, yet sensitive enough to pick up very early signs of age-related pathology. Only a larger cohort study with longitudinal follow-up to ascertain the development of such pathologies would allow us to test this hypothesis. Our test of What-Where-When binding is simple to administer and does not require any special equipment (e.g. virtual reality suite or even a computer), making it more user friendly, especially with older people. The task will need a bit more development to make it more sensitive (larger dynamic range) and will need to be tested with identified patient populations, but we believe it shows promise as a simple and ecologically valid screening task for every-day episodic memory problems.

## 5. Conflicts of interest statement

The authors declare that there are no conflicts of interest.

## 6. Authors and contributors

AM collected the data and did the first analyses and drafting of the manuscript. MB analyzed the order data of the WWW task. JCAR developed the software for the Bayes Factor analysis. PG and TVS developed the original concept and design of the study, trained AM and MB in the relevant tasks and analyses, and further analyzed the data. All authors contributed to the writing of the final manuscript and commented on drafts. All authors accept responsibility for the final manuscript and the results described therein.

## 7. Acknowledgements

AM and MB performed this work as part of their masters degrees at (respectively) the University of Rennes 1 and Newcastle University. The study was funded by Newcastle University.

## 8. References

Aslan, A., and Bäuml, K.-H.T. (2013). Listwise directed forgetting is present in young-old adults, but is absent in old-old adults. *Psychology and Aging* 28**,** 213-218.

Babb, S.J., and Crystal, J.D. (2006). Episodic-like Memory in the Rat. *Current Biology* 16**,** 1317-1321.

Bäckman, L., Small, B.J., and Fratiglioni, L. (2001). Stability of the preclinical episodic memory deficit in Alzheimer's disease. *Brain* 124**,** 96-102.

Baddeley, A.D., Emslie, H., and Nimmo-Smith, I. (1992). "The Speed and Capacity of Language Processing (SCOLP) test.". (Bury St Edmunds, Suffolk: Thames Valley Test Company).

Blachstein, H., Greenstein, Y., and Vakil, E. (2012). Aging and temporal order memory: A comparison of direct and indirect measures. *Journal of Clinical and Experimental Neuropsychology* 34**,** 107-112.

Cheke, L.G., and Clayton, N.S. (2013). Do different tests of episodic memory produce consistent results in human adults? *Learning & Memory* 20**,** 491-498.

Clayton, N.S., and Dickinson, A. (1998). Episodic-like memory during cache recovery by scrub jays. *Nature* 395**,** 272-274.

Crook, T.H., Feher, E.P., and Larrabee, G.J. (1992). Assessment of Memory Complaint in Age- Associated Memory Impairment : The MAC-Q. . *International psychogeriatrics* 4**,** 165-176.

Della Sala, S., Gray, C., Baddeley, A., and Wilson, L. (1997). "The Visual Patterns Test: a new test of short-term visual recall.". (Feltham, Suffolk: Thames Valley Test Company).

Dienes, Z. (2011). Bayesian Versus Orthodox Statistics: Which Side Are You On? *Perspectives on Psychological Science* 6**,** 274-290.

Eacott, M.J., Easton, A., and Zinkivskay, A. (2005). Recollection in an episodic-like memory task in the rat. *Learning and Memory* 12**,** 221-223.

Feeney, M.C., Roberts, W.A., and Sherry, D.F. (2009). Memory for what, where, and when in the black-capped chickadee (*Poecile atricapillus*). *Animal Cognition* 12**,** 767-777.

Gould, K.L., Ort, A.J., and Kamil, A.C. (2012). Do Clark's nutcrackers demonstrate what-where-when memory on a cache-recovery task? *Animal Cognition* 15**,** 37-44.

Greenberg, S.A., and Kurlowicz, L. (2007). The Geriatric Depression Scale ( GDS ) Geriatric Depression Scale : Short Form. *New York university college of nursing* 4.

Hanley, J.R. (1997). Does Articulatory Suppression Remove the Irrelevant Speech Effect? *Memory* 5**,** 423 - 431.

Harris, M.E., Ivnik, R.J., and Smith, G.E. (2002). Mayo's Older Americans Normative Studies: expanded AVLT Recognition Trial norms for ages 57 to 98. *Journal of Clinical and Experimental Neuropsychology* 24**,** 214-220.

Hayne, H., and Imuta, K. (2011). Episodic memory in 3- and 4-year-old children. *Developmental Psychobiology* 53**,** 317-322.

Hedden, T., and Gabrieli, J.D.E. (2004). Insights into the ageing mind: A view from cognitive neuroscience. *Nature Reviews Neuroscience* 5**,** 87-U12.

Holland, S.M., and Smulders, T.V. (2011). Do humans use episodic memory to solve a What-Where-When memory task? *Animal Cognition* 14**,** 95-102.

Inostroza, M., Brotons-Mas, J.R., Laurent, F., Cid, E., and De La Prida, L.M. (2013). Specific Impairment of “What-Where-When” Episodic-Like Memory in Experimental Models of Temporal Lobe Epilepsy. *The Journal of Neuroscience* 33**,** 17749-17762.

Irish, M., Lawlor, B.A., Coen, R.F., and O'mara, S.M. (2011). Everyday episodic memory in amnestic mild cognitive impairment: A preliminary investigation. *BMC Neuroscience* 12.

Jeffreys, H. (1961). *The theory of probability, 3rd Ed.* Oxford, England: Oxford University Press.

Jorm, A.F., Masaki, K.H., Petrovitch, H., Ross, G.W., and White, L.R. (2005). Cognitive deficits 3 to 6 years before dementia onset in a population sample: The Honolulu-Asia Aging Study. *Journal of the American Geriatrics Society* 53, 452-455.

Kessels, R.P.C., Hobbel, D., and Postma, A. (2007). Aging, context memory and binding: a comparison of "What, Where and When" in young and older adults. *International Journal of Neuroscience* 117, 795-810.

Kessels, R.P.C., Postma, A., and De Haan, E.H.F. (1999). Object Relocation: A program for setting up, running, and analyzing experiments on memory for object locations. *Behavior Research Methods, Instruments, & Computers* 31, 423–428.

Lezak, M.D., Howieson, D.B., and Loring, D.W. (2004). *Neuropsychological assessment.* New York: Oxford University Press.

Lundervold, A.J., Wollschlager, D., and Wehling, E. (2014). Age and sex related changes in episodic memory function in middle aged and older adults. *Scandinavian Journal of Psychology* 55, 225-232.

Mammarella, N., Borella, E., Carretti, B., Leonardi, G., and Fairfield, B. (2013). Examining an emotion enhancement effect in working memory: Evidence from age-related differences. *Neuropsychological Rehabilitation* 23, 416-428.

Marks, D. (1973). Visual Imagery differences in the recall of pictures. *British Journal of Psychology* 64, 17-24.

Masur, D.M., Sliwinski, M., Lipton, R.B., Blau, A.D., and Crystal, H.A. (1994). Neuropsychological Prediction of Dementia and the Absence of Dementia in Healthy Elderly Persons. *Neurology* 44, 1427-1432.

Newcombe, N.S., Balcomb, F., Ferrara, K., Hansen, M., and Koski, J. (2014). Two rooms, two representations? Episodic-like memory in toddlers and preschoolers. *Developmental Science* 17, 743-756.

Nyberg, L., Lovden, M., Riklund, K., Lindenberger, U., and Backman, L. (2012). Memory aging and brain maintenance. *Trends in Cognitive Sciences* 16, 292-305.

Plancher, G., Gyselinck, V., Nicolas, S., and Piolino, P. (2010). Age effect on components of episodic memory and feature binding: A virtual reality study. *Neuropsychology* 24, 379-390.

Plancher, G., Tirard, A., Gyselinck, V., Nicolas, S., and Piolino, P. (2012). Using virtual reality to characterize episodic memory profiles in amnestic mild cognitive impairment and Alzheimer's disease: Influence of active and passive encoding. *Neuropsychologia* 50, 592-602.

Rey, A. (1964). *L'examen Clinique en psychologie.* Paris: Presse universitaire de France.

Ricci, M., Graef, S., Blundo, C., and Miller, L.A. (2012). Using the Rey Auditory Verbal Learning Test (RAVLT) to Differentiate Alzheimer's Dementia and Behavioural Variant Fronto-Temporal Dementia. *Clinical Neuropsychologist* 26, 926-941.

Roberts, W.A., Feeney, M.C., Macpherson, K., Petter, M., Mcmillan, N., and Musolino, E. (2008). Episodic-Like Memory in Rats: Is It Based on When or How Long Ago? *Science* 320**,** 113-115.

Russell, J., Cheke, L.G., Clayton, N.S., and Meltzoff, A.N. (2011). What can What–When–Where (WWW) binding tasks tell us about young children's episodic foresight? Theory and two experiments. *Cognitive Development* 26**,** 356-370.

Russell, J., and Hanna, R. (2012). A Minimalist Approach to the Development of Episodic Memory. *Mind & Language* 27**,** 29-54.

Sbordone, R.J., and Long, C.J. (1996). *Ecological validity of neuropsychological testing.* . Delray Beach, FL: St. Lucie Press.

Suddendorf, T., and Corballis, M.C. (1997). Mental time travel and the evolution of the human mind. *Genetic Social and General Psychology Monographs* 123**,** 133-167.

Sunderland, A., Harris, J.E., and Baddeley, A.D. (1983). Do laboratory tests predict everyday memory? A neuropsychological study. *Journal of Verbal Learning and Verbal Behaviour* 22**,** 727–738.

Tierney, M.C., Nores, A., Snow, W.G., Fisher, R.H., Zorzitto, M.L., and Reid, D.W. (1994). Use of the Rey Auditory Verbal Learning Test in differentiating normal aging from Alzheimer's and Parkinson's dementia. *Psychological Assessment* 6**,** 129-134.

Vakil, E., and Blachstein, H. (1994). A supplementary measure in the Rey AVLT for assessing incidental learning of temporal order. *Journal of Clinical Psychology* 50**,** 240-245.

Wechsler, D. (1981). WAIS-R manual, Wechsler Adult Intelligence Scale-Revised. Cleveland, OH: Psychological Corp.

Yi, Y., and Friedman, D. (2013). Age-related differences in working memory: ERPs reveal age-related delays in selection- and inhibition-related processes. *Aging, Neuropsychology, and Cognition* 21**,** 483-513.

Zavagnin, M., Borella, E., and De Beni, R. (2014). When the mind wanders: Age-related differences between young and older adults. *Acta Psychologica* 145**,** 54-64.

Zinkivskay, A., Nazir, F., and Smulders, T.V. (2009). What-Where-When memory in magpies (*Pica pica*). *Animal Cognition* 12**,** 119-125.

## 9. Figure Captions

**Figure 1.** Comparison of the three age groups on **A.** the Everyday Memory Questionnaire (EMQ); **B.** the Memory Complaint Questionnaire (MAC-Q); **C.** verbal (Forward Digit Span, Backward Digit Span) and visual (Spatial Span, Visual Patterns, CANTAB PA) working memory tasks; and **D.** psychomotor speed (Trail Making A), Executive Function (Trail Making B) and vocabulary (SCOLP tasks: W=Words; Cp= Sentence Comprehension). We plotted means + SEM; * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

**Figure 2.** Comparison of the three age groups in their performance on different measures of the Rey AVLT. **A.** Number of words (out of 15) recalled in the different phases of the task. RI=Retroactive Interference (A5-A6); PI=Proactive Interference (A1-B). Note the very similar trajectories for the two older groups. **B.** d-prime score on the recognition task. **C., D., E.** Different measures of the memory for the order of the words in the list: **C.** the number of words that were placed in their correct position (hits); **D.** the sum of the absolute differences between the original position and the remembered position of each word in the list; **E.** the Pearson's correlation coefficients between the original order and the remembered order. We plotted means ± SEM; * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

**Figure 3.** Comparison of the three age groups in their performance on the Kessels Object-Location Binding task. **A.** Visuospatial reconstruction. The error score is the sum of the distance (in mm) between the original and reconstructed locations of the objects. **B.** Position Only Memory. The error score is the sum of the distance between the remembered locations and the closest original locations of the objects. This score is statistically controlled for the error score on the VSR (see Methods). **C.** Combined Memory Score. The error score is the sum of the distance (in mm) between the original and remembered locations of the objects. This score is statistically controlled for the error score on the VSR (see Methods). **D.** Object Recognition Memory. The error score is the number of incorrectly identified objects (out of 10). **E.** Object-Location Binding. The error score is the number of marked locations with an incorrect object assigned to them (out of 10). We plotted means + SEM; * $p<0.05$, *** $p<0.001$.

**Figure 4.** Comparison of the three age groups on the WWW memory task. The graph represents the proportion of correct objects in each of the categories, excluding all other categories (see Methods). For example, proportion of correct What-Where combinations is out of the total number of objects that have not been remembered in a complete WWW combination, and the proportion of correct locations (Where) is out of the number of locations that have not been remembered in any combination at all. We plotted means + SEM; ** $p<0.01$

**Figure 5.** Comparison of the three age groups on: **A.** the average vividness score, split by those participants who claimed to re-experience the event (Remember) and those who just knew the information (Know). The numbers on the bars represent the number of individuals in each condition. **B.** The mean number of WWW combinations recalled by participants, split in the same manner as in A. We plotted means + SEM.

**Figure 6.** Regression plots of performance on the WWW binding (number of correct combinations out of 16) as predicted by: **A.** the number of words remembered after one reading of the list in the RAVLT (A1); **B.** the number of words forgotten over the 30-min retention interval in the RAVLT (A6-A7; negative numbers indicate more correct words at A7 than at A6); **C.** the number of words from list A forgotten while learning and repeating list B (A5-A6; Retroactive Interference; negative numbers indicate more correct words at A7 than at A6); **D.** Error score on the Combined Object-Location Memory (note that one 18-25 and one 70+ participant had missing data for this task); **E.** the Memory Complaint Questionnaire (MAC-Q; higher scores indicate more complaints); **F.** the Everyday Memory Questionnaire (EMQ; higher scores indicate more problems). Continuous lines and filled circles: 18-25; long dashes and open circles: 60-69; short dashes and closed triangles: 70+. Significance levels indicated in the panels are for the overall effect of the predictor on the WWW performance. For more details of the analyses, see the main text.

Figure 1.TIF

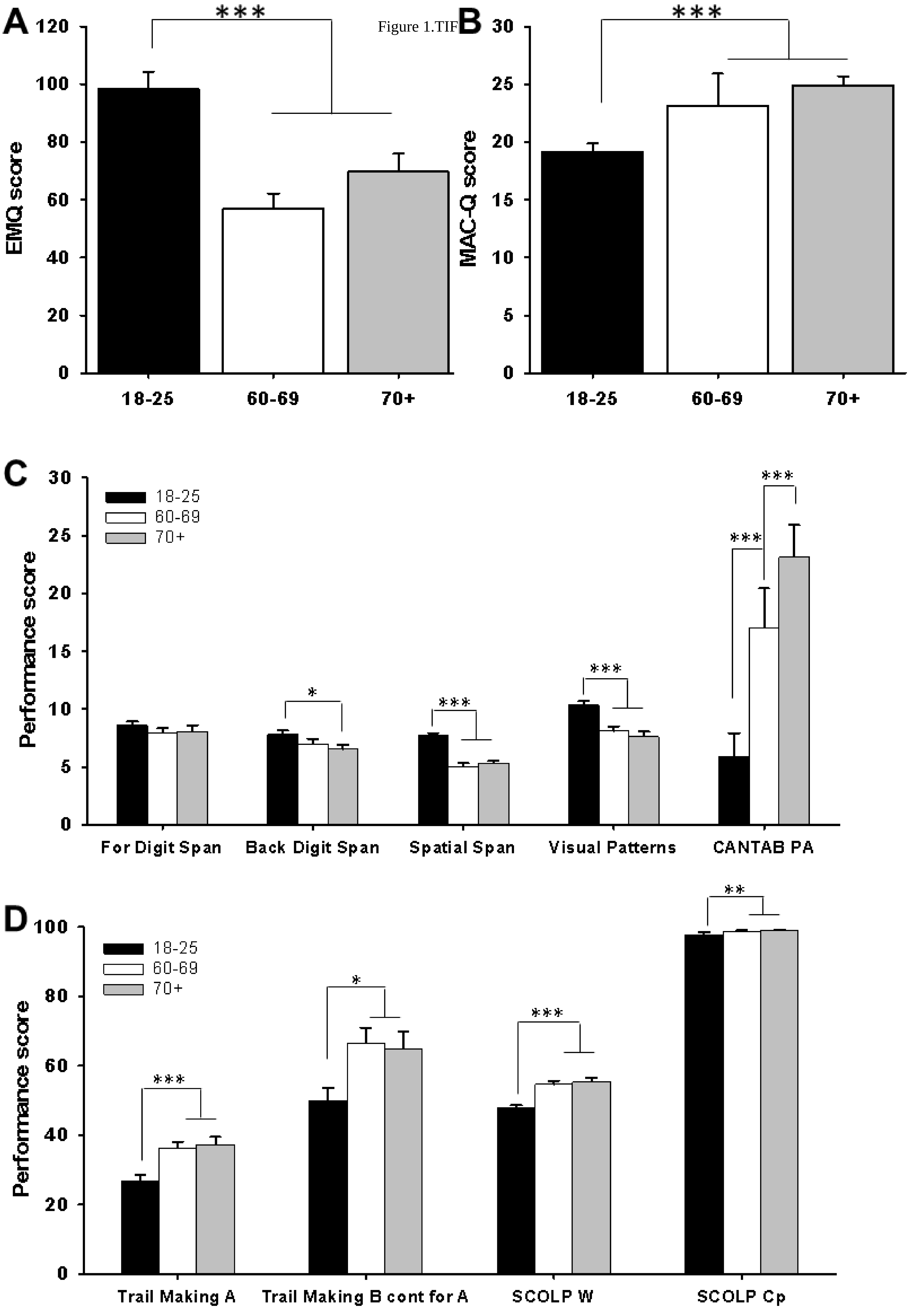

A
B
C
D
EMQ score
MAC-Q score
Performance score
18-25
60-69
70+
For Digit Span
Back Digit Span
Spatial Span
Visual Patterns
CANTAB PA
Trail Making A
Trail Making B cont for A
SCOLP W
SCOLP Cp
***
**
*

Figure 2.TIF

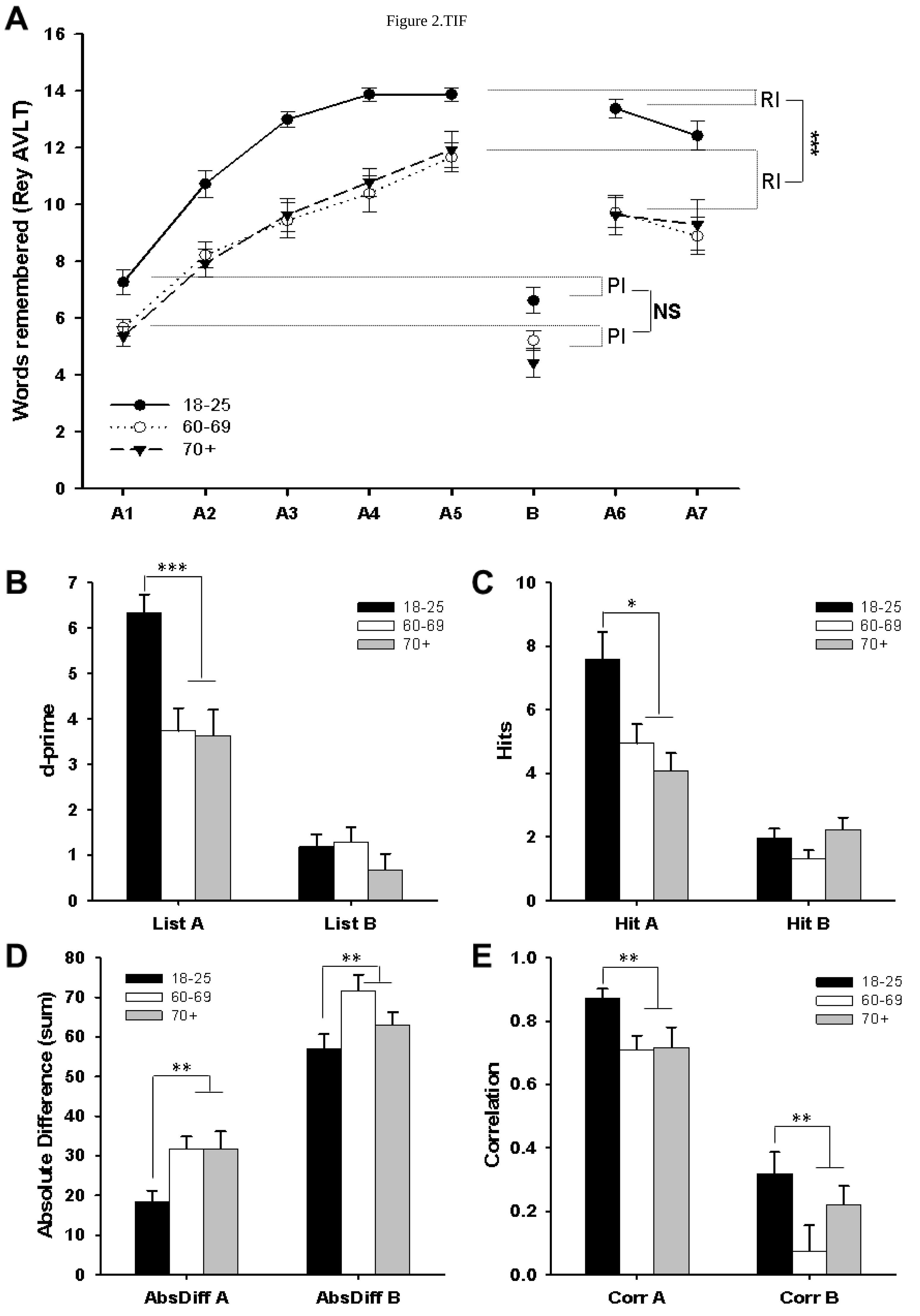

A
Words remembered (Rey AVLT)
16
14
12
10
8
6
4
2
0
A1
A2
A3
A4
A5
B
A6
A7
RI
RI
***
PI
PI
NS
18-25
60-69
70+
B
d-prime
***
List A
List B
C
Hits
*
Hit A
Hit B
D
Absolute Difference (sum)
**
**
AbsDiff A
AbsDiff B
E
Correlation
**
**
Corr A
Corr B

Figure 3.TIF

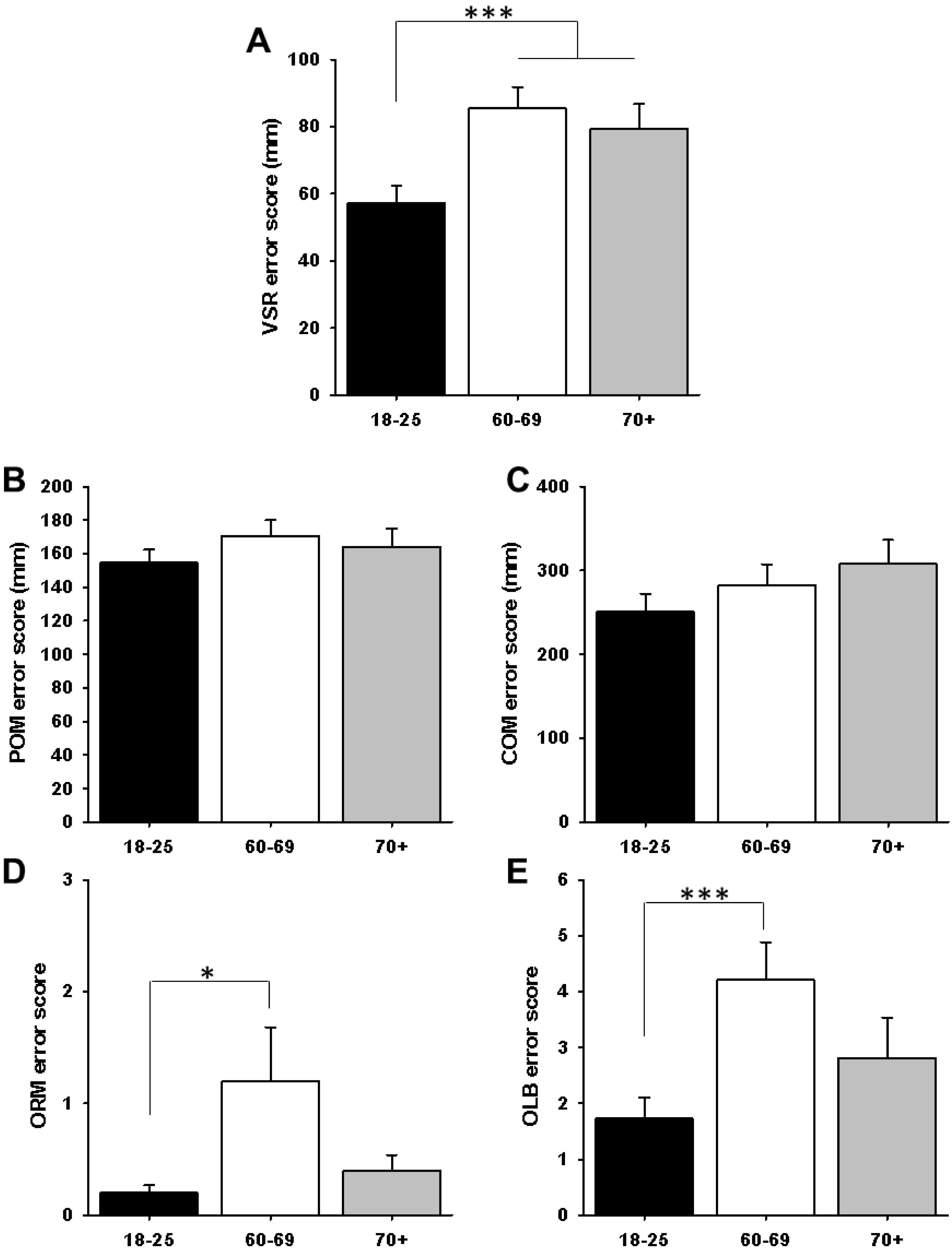
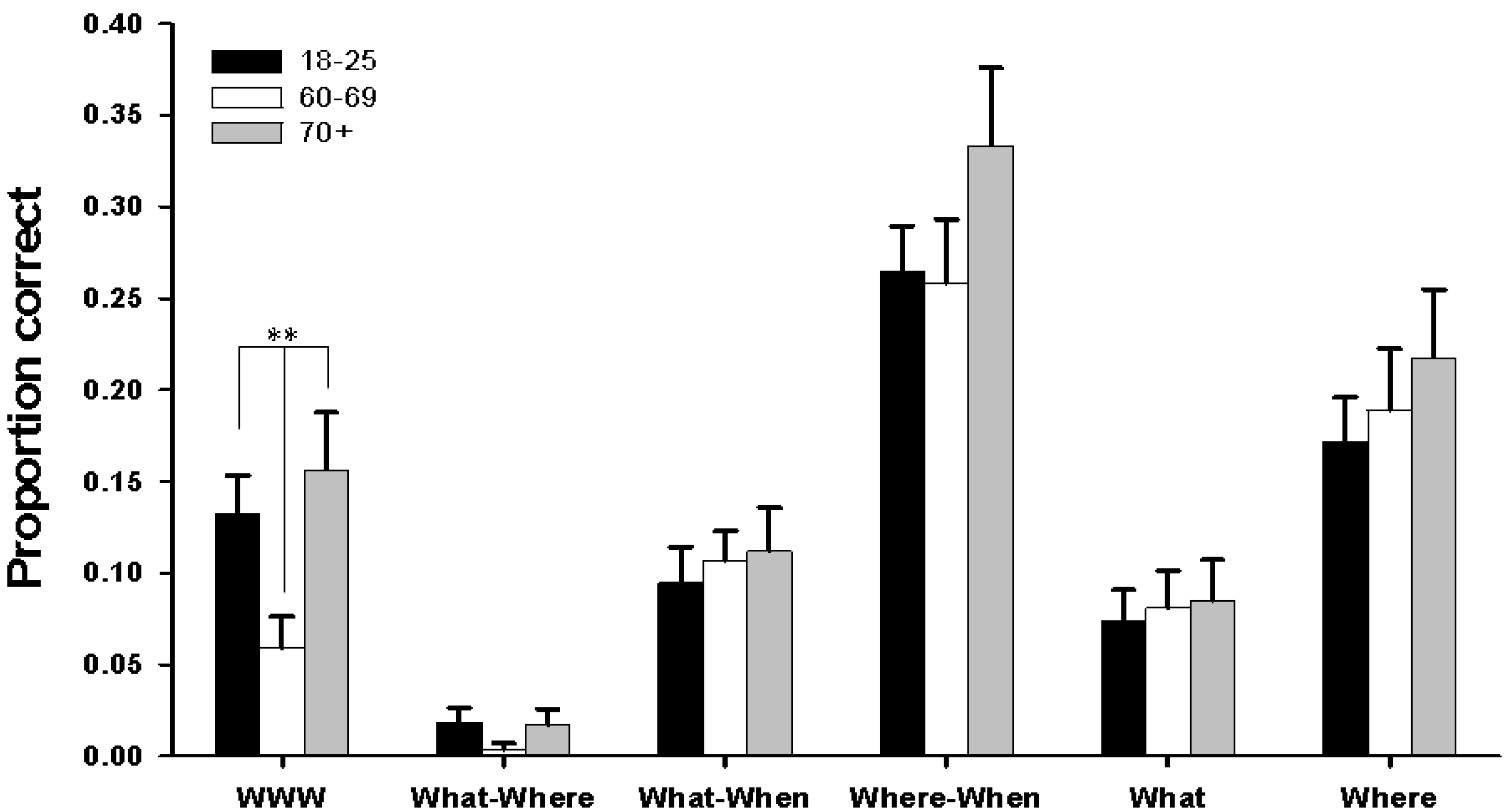

A
***
VSR error score (mm)
0
20
40
60
80
100
18-25
60-69
70+
B
POM error score (mm)
0
20
40
60
80
100
120
140
160
180
200
18-25
60-69
70+
C
COM error score (mm)
0
100
200
300
400
18-25
60-69
70+
D
*
ORM error score
0
1
2
3
18-25
60-69
70+
E
***
OLB error score
0
1
2
3
4
5
6
18-25
60-69
70+

Figure 4.TIF



Proportion correct
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
18-25
60-69
70+
**
WWW
What-Where
What-When
Where-When
What
Where

Figure 5.TIF

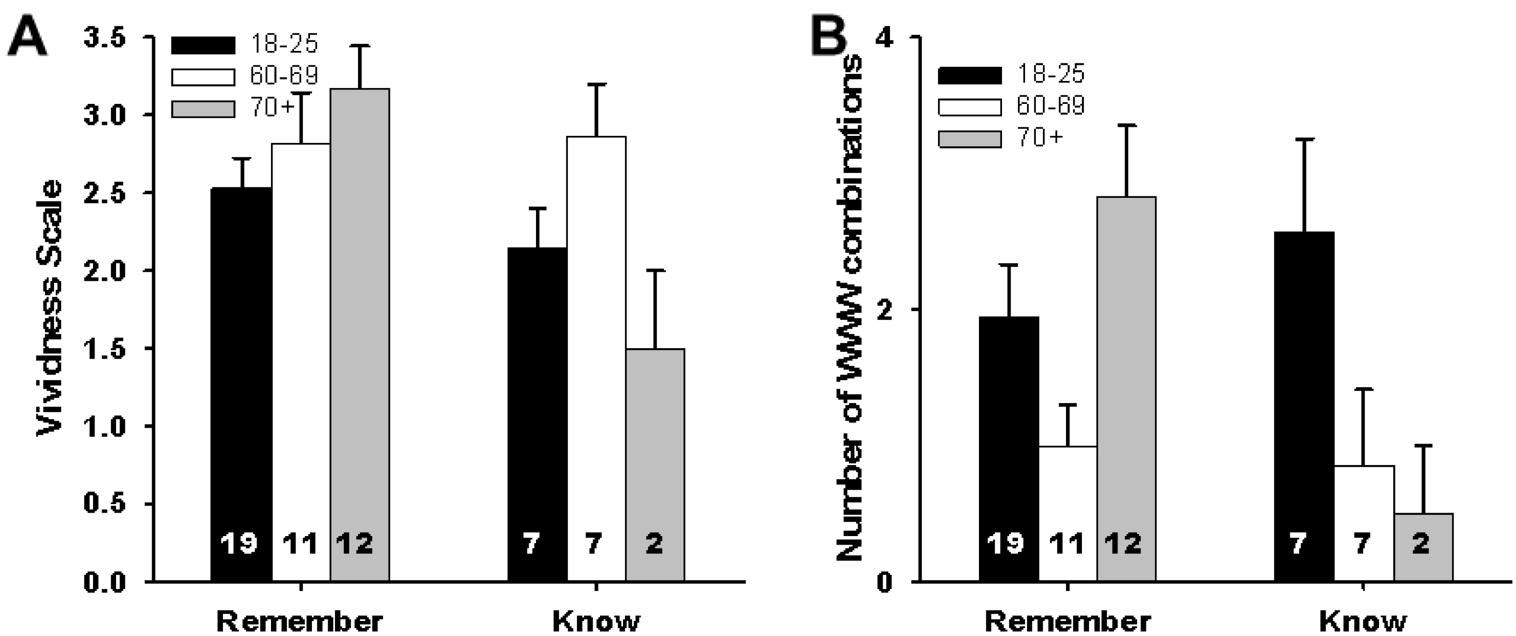
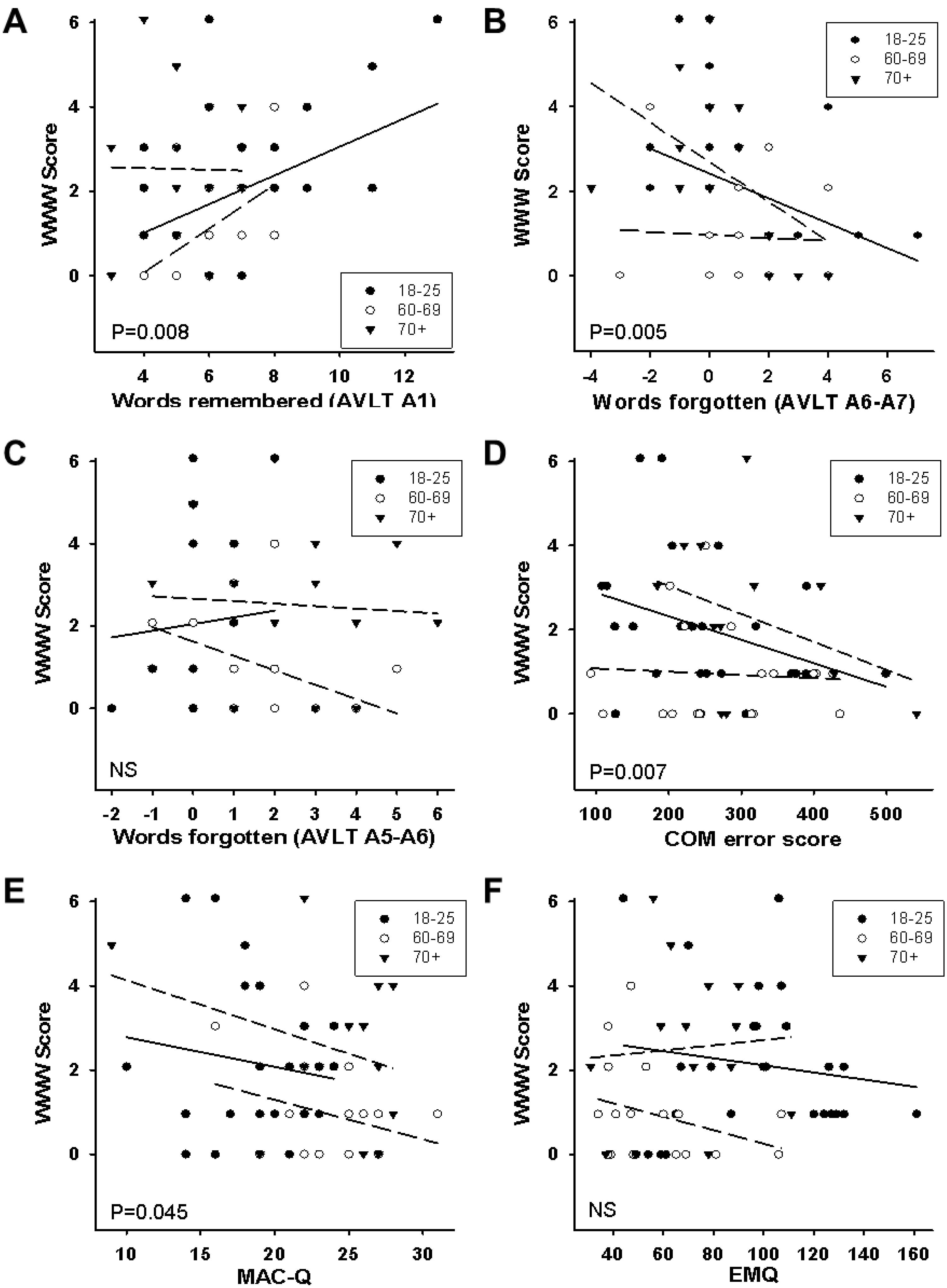

A
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Vividness Scale
18-25
60-69
70+
19
11
12
7
7
2
Remember
Know
B
4
2
0
Number of WWW combinations
18-25
60-69
70+
19
11
12
7
7
2
Remember
Know

Figure 6.TIF


A
WWW Score
P=0.008
18-25
60-69
70+
Words remembered (AVLT A1)
B
WWW Score
P=0.005
18-25
60-69
70+
Words forgotten (AVLT A6-A7)
C
WWW Score
NS
18-25
60-69
70+
Words forgotten (AVLT A5-A6)
D
WWW Score
P=0.007
18-25
60-69
70+
COM error score
E
WWW Score
P=0.045
18-25
60-69
70+
MAC-Q
F
WWW Score
NS
18-25
60-69
70+
EMQ