# Changes in grey matter induced by training

Newly honed juggling skills show up as a transient feature on a brain-imaging scan.

Does the structure of an adult human brain alter in response to environmental demands[1,2]? Here we use whole-brain magnetic-resonance imaging to visualize learning-induced plasticity in the brains of volunteers who have learned to juggle. We find that these individuals show a transient and selective structural change in brain areas that are associated with the processing and storage of complex visual motion. This discovery of a stimulus-dependent alteration in the brain's macroscopic structure contradicts the traditionally held view that cortical plasticity is associated with functional rather than anatomical changes.

Animal studies indicate that experience-related changes may occur in mammalian brain structures, but so far there has been no evidence of comparable alterations in the human brain[3–5]. To investigate this possibility, we divided a homogeneous group of volunteers (21 female, 3 male; mean age, 22 yr ± 1.6 s.d.), who were matched for sex and age, into two groups, designated as jugglers and non-jugglers. Both groups were inexperienced in juggling at the time of their first brain scan.

Subjects in the juggler group were given 3 months to learn a classic three-ball cascade juggling routine. A second brain scan was performed when they had become skilled performers (that is, when they could sustain juggling for at least 60 seconds). A third scan was carried out 3 months later; during the intervening period, none of the jugglers practised or attempted to extend their skills — for example, by learning a four-ball or a reverse cascade. In fact, most subjects were no longer fluent in three-ball cascade juggling by the time of the third scan.

We used voxel-based morphometry, a sophisticated objective whole-brain technique, to investigate subtle, region-specific changes in grey and white matter by averaging results across the volunteers. This method is based on high-resolution, three-dimensional magnetic-resonance imaging, registered in a common stereotactic space, and is designed to find significant regional differences by applying voxel-wise statistics in the context of gaussian random fields[6,7].

Group comparison at the beginning (the baseline) showed no significant regional differences in grey matter between jugglers and non-jugglers. In the longitudinal analysis, the juggler group demonstrated a significant (44 d.f., $P<0.05$) transient bilateral expansion in grey matter in the mid-temporal area (hMT/V5) and in the left posterior intraparietal sulcus between the first and the second scans. This expansion decreased in the third scan (Fig. 1). We found a close relationship in these regions between the transient structural grey-matter changes and the juggling performance. These findings were specific to the training stimulus, as the non-jugglers showed no change in grey matter over the same period.

THE IMAGE BANK / GETTY

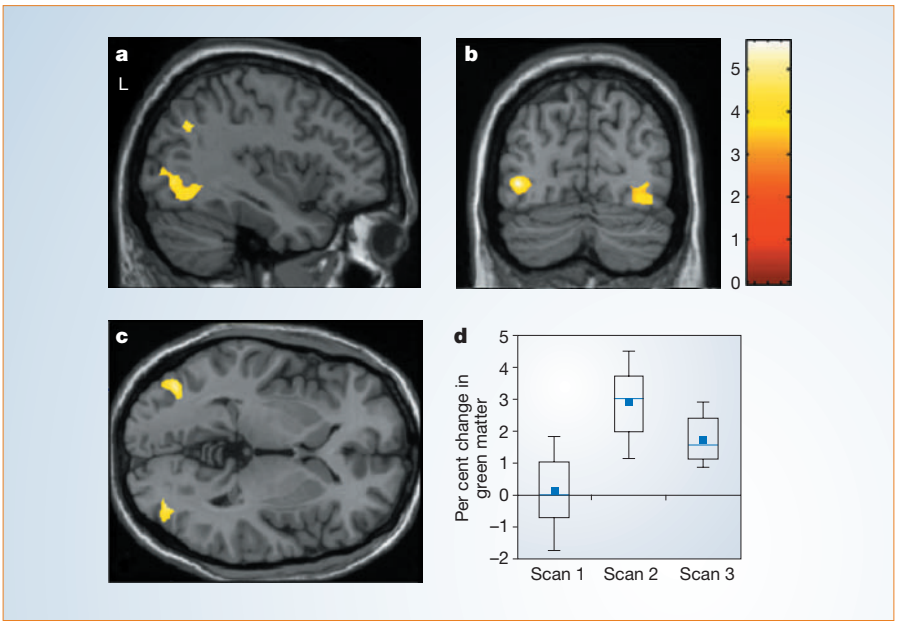


**Figure 1** Transient changes in brain structure induced while learning to juggle. **a–c,** Statistical parametric maps showing the areas with transient structural changes in grey matter for the jugglers group compared with non-juggler controls. **a,** Sagittal view; **b,** coronal view; **c,** axial view. The increase in grey matter is shown superimposed on a normalized T1 image. The left side (L) of the brain is indicated. A significant expansion in grey matter was found between the first and second scans in the mid-temporal area (hMT/V5) bilaterally (left: *x*, –43; *y*, –75; *z*, –2, with $Z = 4.70$; right: *x*, 33; *y*, –82; *z*, –4, with $Z = 4.09$) and in the left posterior intraparietal sulcus (*x*, –40; *y*, –66; *z*, 43 with $Z = 4.57$), which had decreased by the time of the third scan. Colour scale indicates *Z* scores, which correlate with the significance of the change. **d,** Relative grey-matter change in the peak voxel in the left hMT for all jugglers over the three time points. The box plot shows the standard deviation, range and the mean for each time point.

Our results contradict the traditionally held view that the anatomical structure of the adult human brain does not alter, except for changes in morphology caused by ageing or pathological conditions. Our findings indicate that learning-induced cortical plasticity is also reflected at a structural level.

As all of our volunteers have normal fine-motor skills, we conclude that juggling, and consequently the perception and spatial anticipation of moving objects, is a stronger stimulus for structural plasticity in the visual areas (used for the retention of visual-motion information[8,9]) than in the motor areas (involved in the planning and execution of coordinate motion — that is, the supplementary motor area and/or the motor cortex, cerebellum and basal ganglia).

Although the observed transient increase in grey matter takes place in specific motion-selective areas, the microscopic changes underlying these dynamic structural alterations are unclear. Macroscopic alterations may be based on changes at the level of

synaptic bulk and neurites, or they might include increased cell genesis, for example, of glial or even neuronal cells[4]. Imaging results need to be compared with histological data for identification of the structural basis at the microscopic level of temporary, training-dependent structural changes in our brains.

**Bogdan Draganski*, Christian Gaser†, Volker Busch*, Gerhard Schuierer‡, Ulrich Bogdahn*, Arne May***

**Department of Neurology, and ‡Institute of Neuroradiology, University of Regensburg, Regensburg 93053, Germany*
*e-mail: arne.may@klinik.uni-regensburg.de*
*†Department of Psychiatry, University of Jena, 07740 Jena , Germany*

1. Draganski, B. *et al. Nature Med.* **8,** 1186–1188 (2002).
2. Maguire, E. A. *et al. Proc. Natl Acad. Sci. USA* **97,** 4398–4403 (2000).
3. Kempermann, G., Gast, D. & Gage, F. H. *Ann. Neurol.* **52,** 135–143 (2002).
4. Trachtenberg, J. T. *et al. Nature* **420,** 788–794 (2002).
5. Grutzendler, J., Kasthuri, N. & Gan, W. B. *Nature* **420,** 812–816 (2002).
6. Ashburner, J. & Friston, K. J. *Neuroimage* **11,** 805–821 (2000).
7. Good, C. D. *et al. Neuroimage* **17,** 29–46 (2002).
8. Bisley, J. W. & Pasternak, T. *Cereb. Cortex* **10,** 1053–1065 (2000).
9. Sereno, M. I., Pitzalis, S. & Martinez, A. *Science* **294,** 1350–1354 (2001).

**Competing financial interests**: declared none.

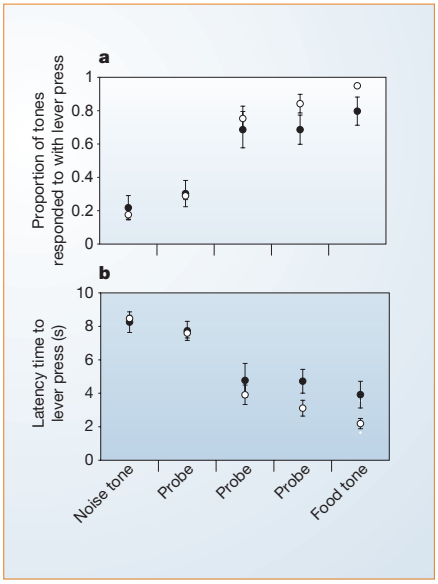


**Figure 1** Mean (±1 s.e.) responses to tones during 10 daily 30-min test sessions for male Lister hooded rats housed under 'predictable' (open circles, $n=4$) and 'unpredictable' (filled circles, $n=5$) conditions. **a,** Proportion of tones to which the animals responded to by pressing a lever. **b,** Latent time between sounding of the tone and pressing of the lever. 'Noise' and 'food' tones are the tones used during training (2 and 4 kHz, respectively, for about half of the rats, and 4 and 2 kHz, respectively, for the remaining rats). 'Probe' tones are non-reinforced, intermediate tones (2.5, 3, 3.5 kHz), each randomly interspersed with a probability of 0.085 between the reinforced training tones. Regression equations were calculated for each rat, correcting for nonlinear relationships by using binary logistic regression (for proportions) and logarithmic transformations for linear regression (for latencies). Animals were checked daily and remained healthy throughout the experiments.

Animal behaviour

# Cognitive bias and affective state

Information processing by humans can be biased by their emotions — for example, anxious and depressed people tend to make negative judgements about events and to interpret ambiguous stimuli unfavourably[1–4]. Here we show that such a 'pessimistic' response bias can also be measured in rats that are housed in unpredictable conditions[5,6]. Our findings indicate that cognitive bias can be used as an indicator of affective state in animals, which should facilitate progress in animal-welfare studies.

We trained rats to respond by pressing a lever when they heard a tone associated with a positive event (delivery of a 45-mg food pellet) and to refrain from pressing the lever as a way to avoid a negative event (an unpleasant burst of white noise) when they heard another tone. Once the animals were able to score a correct response to each tone more than 50% of the time (binomial testing for three consecutive daily 30-min sessions), they were allocated to either 'unpredictable' housing, which induces symptoms of a mild depression-like state[5,6], or to 'predictable' housing.

In 'unpredictable' housing, between zero and two negative interventions were made at random times on any one day — for example, the cage might be unfamiliar or tilted, or it could contain a stranger of the same species; sometimes the light/dark cycle would be temporarily reversed or bedding left damp. These changes were never imposed simultaneously, and they were made at least two hours before or after test sessions. 'Predictable' housing, in contrast, was maintained as during training, with none of these interventions.

After nine days, during which training was continued, the rats were exposed to non-reinforced tones that had frequencies intermediate between those of the two food-delivery and noise-avoidance tones. Ten test sessions were held to investigate the animals' anticipation of these positive or negative events, as judged by their lever-press response to these ambiguous tones.

The proportion of tones responded to by lever pressing (Fig. 1a) and the time taken to respond to the tones (mean response latencies; Fig. 1b) were calculated for each tone for each rat on each of the test days. Analysis of variance with repeated-measures (tone, test day) and a between-subjects factor (housing) revealed a housing × tone interaction ($F_{4,28}=2.72, P<0.05$).

The proportion of tones responded to with a lever press by individuals kept in unpredictable housing indicated that fewer lever presses were made in response to tones of frequency close to that of the food tone (Fig. 1a) (in two-tailed $t$-tests, $t=1.88$, d.f. $=7$, $P=0.1$). These rats were also slower to press the lever in response to the food tone and to the ambiguous tones that were close to it in frequency (Fig. 1b) ($t=-2.44$, d.f. $=7$, $P<0.05$). Both findings were still valid when only the responses by the rats to the ambiguous tones were analysed (proportions: $t=1.92$, d.f. $=7$, $P=0.09$; latencies: $t=-2.42$, d.f. $=7$, $P<0.05$).

Overall, rats in unpredictable housing were slower to respond and tended to show fewer responses to ambiguous tones close to the positive tone and to this tone itself. The treatment groups did not differ ($P>0.2$) in tests of feeding motivation (consumption speed of freely available food pellets[7]), anhedonia (amount of sucrose solution consumed[5,6]), activity (hole–board test[8]), bodyweight change across the test period, and response accuracy to training tones before and after the imposition of housing changes, indicating that none of these factors was likely to account for our findings.

By using ambiguous stimuli to probe animals' relative anticipation of positive and negative events, we have shown that rats in unpredictable housing show behaviour indicating reduced anticipation of a positive event. This compares with findings for depressed or anxious humans, who also have reduced expectation of positive events[1,4] and interpret ambiguous stimuli negatively[3].

Our results call for further investigation of the underlying processes involved[9,10]. We find no evidence of enhanced anticipation of the negative event. This may be due to a floor effect and could be revealed using, for example, lever-pressing and nose-poking as counterbalanced positive and negative responses. It is possible that our technique could be adapted to detect an enhanced expectation of positive events — a correlate of happy mood in humans[4]. Being able to assess positive as well as negative affect in animals is an important objective for animal welfare[11].

**Emma J. Harding, Elizabeth S. Paul, Michael Mendl**

*Centre for Behavioural Biology, Department of Clinical Veterinary Science, University of Bristol, Langford House, Langford BS40 5DU, UK*
*e-mail: mike.mendl@bris.ac.uk*

1. MacLeod, A. K. & Byrne, A. J. *Abnorm. Psychol.* **105,** 286–289 (1996).
2. Gotlib, I. H. & Krasnoperova, E. *Behav. Therapy* **29,** 603–617 (1998).
3. Eysenck, M. W. *et al. J. Abnorm. Psychol.* **100,** 144–150 (1991).
4. Wright, W. F. & Bower, G. H. *Organiz. Behav. Hum. Decis. Process* **62,** 276–291 (1992).
5. Willner, P. *Psychopharmacology* **134,** 319–329 (1997).
6. Zurita, A. *et al. Behav. Brain Res.* **117,** 163–171 (2000).
7. Nielsen, B. L. *Appl. Anim. Behav. Sci.* **63,** 79–91 (1999).
8. Fernandes, C. & File, S. E. *Pharmacol. Biochem. Behav.* **54,** 31–40 (1996).
9. Spruijt, B. M., van den Bos, R. & Pijlman, F. T. A. *Appl. Anim. Behav. Sci.* **72,** 145–171 (2001).
10. Berridge, K. C. & Robinson, T. E. *Trends Neurosci.* **26,** 507–513 (2003).
11. Dawkins, M. S. in *Coping with Challenge. Welfare in Animals including Humans* (ed. Broom, D.M.) 63–76 (Dahlem University Press, Berlin, 2001).

**Competing financial interests**: declared none.

Brain and Cognition 90 (2014) 174–180

Contents lists available at ScienceDirect

Brain and Cognition

journal homepage: www.elsevier.com/locate/b&c

# The effects of musical practice on structural plasticity: The dynamics of grey matter changes

Mathilde Groussard [a,b,c,d,*], Fausto Viader [a,b,c,e], Brigitte Landeau [a,b,c,d], Béatrice Desgranges [a,b,c,d], Francis Eustache [a,b,c,d], Hervé Platel [a,b,c,d]

[a] INSERM, U1077, Caen, France
[b] Université de Caen Basse-Normandie, UMR-S1077, Caen, France
[c] Ecole Pratique des Hautes Etudes, UMR-S1077, Caen, France
[d] CHU de Caen, U1077, Caen, France
[e] CHU de Caen, Service de Neurologie, Caen, France

ARTICLE INFO

*Article history:*
Accepted 23 June 2014
Available online 13 August 2014

*Keywords:*
Music
Training
Plasticity
MRI
VBM

ABSTRACT

Intensive training and the acquisition of expertise are known to bring about structural changes in the brain. Musical training is a particularly interesting model. Previous studies have reported structural brain modifications in the auditory, motor and visuospatial areas of musicians compared with nonmusicians. The main goal of the present study was to go one step further, by exploring the dynamic of those structural brain changes related to musical experience. To this end, we conducted a regression study on 44 nonmusicians and amateur musicians with 0–26 years of musical practice of a variety instruments. We sought first to highlight brain areas that increased with the duration of practice and secondly distinguish (thanks to an ANOVA analysis) brain areas that undergo grey matter changes after only limited years of musical practice from those that require longer practice before they exhibit changes. Results revealed that musical training results a greater grey matter volumes in different brain areas for musicians. Changes appear gradually in the left hippocampus and right middle and superior frontal regions, but later also include the right insula and supplementary motor area and left superior temporal, and posterior cingulate areas. Given that all participants had the same age and that we controlled for age and education level, these results cannot be ascribed to normal brain maturation. Instead, they support the notion that musical training could induce dynamic structural changes.



## 1. Introduction

The effects of intensive training and expertise on brain structure have been observed in several areas (for reviews, see Draganski & May, 2008; May, 2010; Zatorre, Fields, & Johansen-Berg, 2012), including the taxi driver's hippocampus (Maguire et al., 2000), the juggler's midtemporal area (hMT/V5) and left posterior intraparietal sulcus (Boyke, Driemeyer, Gaser, Büchel, & May, 2008; Draganski et al., 2004), the basketball player's cerebellum (Park et al., 2009) and the frontal and parietal areas of individuals who engage in physical exercise (Taubert et al., 2010).

It is now well established that musical training requires complex multimodal abilities, including somatosensory and memory processes, motor skills and emotion. Thus, musical expertise can be regarded as a relevant model for studying structural brain plasticity mechanisms (Wan & Schlaug, 2010). Skills acquired by musicians result not only in specific connections and interactions between different brain areas (Altenmüller, 2008; Fauvel et al., 2014), but also in the enlargement of brain regions involved in music-related processes such as auditory, motor and visuospatial abilities (Bermudez, Lerch, Evans, & Zatorre, 2009; Gaser & Schlaug, 2003; James et al., 2014; Luders, Gaser, Jancke, & Schlaug, 2004; Schneider et al., 2002; Seung, Kyong, Woo, Lee, & Lee, 2005). In most studies, structural changes have been observed after only a few months of intensive practice. For example, Hyde et al. (2009) showed that 15 months of instrumental musical training in childhood were enough to increase the volume of the auditory and motor cortices. Moreover, some authors have also shown a relationship between the age of onset of musical training and structural brain modifications particularly in the premotor cortex and corpus callosum (Bailey, Zatorre, & Penhune, 2014; Steele, Bailey, Zatorre, & Penhune, 2013) for early-trained musicians

* Corresponding author. Address: Inserm – EPHE-Université de Caen/Basse-Normandie-CHU de Caen, Unité U1077, Laboratoire de Cyceron, Blvd Becquerel BP 5229, 14074 Caen cedex 5, France. Fax: +33 (0)2 31 47 01 06.
*E-mail address:* groussard@cyceron.fr (M. Groussard).

http://dx.doi.org/10.1016/j.bandc.2014.06.013
0278-2626/

(<age 7) suggesting a sensitive period of musical training effect. Nevertheless, these training-induced, regional structural brain changes do not occur solely during brain development, as they can be observed throughout the lifespan (Engvig et al., 2010). Nor are they restricted to specific cognitive demands (for a review, see Draganski & May, 2008), as they can concern many areas sustaining learning, memory, sensory or motor processes (see Fauvel, Groussard, Eustache, Desgranges, & Platel, 2013 for review in musical training). Thus, some investigations have revealed that musical training can have structural and functional effects on regions not directly involved in sensori-motor processes of music practice, such as those that subtend working memory (Oechslin, Van De Ville, Lazeyras, & James, 2012; Schulze, Mueller, & Koelsch, 2011) or long-term memory (Groussard et al., 2010a). In the latter study, a musical memory task revealed both functional and structural greater activation and grey matter in the left hippocampus of adult musicians compared to nonmusicians. Regarding results of our previous study, we would like to go one step further in the assessment of the relationship between the duration of musical expertise and the left hippocampus and generally in whole-brain grey matter changes.

The main goal of the present study is to analyze in what way structural grey matter changes with increasing number of years of musical training. To this end, we conducted a regression study on grey matter volumes of 44 nonmusicians and amateur musicians with 0–26 years of musical practice of a variety instruments. We sought first to highlight brain areas that increased in volume with the duration of practice and secondly distinguish (thanks to an ANOVA analysis) brain areas that underwent grey matter changes after only limited years of musical practice from those that require longer practice before they exhibit changes. Actually, some neuroimaging studies on structural plasticity (for a review, see Jancke, 2009) have suggested that musical training initially induces structural changes in regions that are directly involved in music learning (i.e. auditory and motor cortices). But recently James et al. (2014) studied the expertise effect on grey matter changes comparing three groups: nonmusicians, amateur musicians and professional musicians and observed that grey matter areas related to higher-order cognitive function increase with musical practice.

We therefore hypothesized that musical practice would have differential effects on the brain according to its duration, affecting first regions involved in motor and perceptual processes, to subsequently include regions involved in higher cognitive processes (i.e. executive functions, memory, and emotion).

## 2. Methods

### 2.1. Participants

Forty-four young volunteers (26 men, mean age ± *SD*: 23.75 ± 3.43 years, mean education level ± *SD*: 15.45 ± 2.02 years) with no history of neurological or psychiatric disease took part in this study. All participants were right-handed according to the Edinburgh Inventory (Oldfield, 1971), none of them reported having hearing deficits and none had perfect pitch. This study was approved by the regional ethics committee, and written informed consent was obtained from all the participants.

In order to study the progression of musical expertise from the very beginning, eleven nonmusician participants were included. They were classified as strictly nonmusicians, and met the following criteria: (1) none had ever taken part in musical performances or received music lessons (except for basic musical education at French high school, corresponding to 1 hour/week), (2) they were *common listeners* (i.e., not music lovers, who tend to listen to one specific type of music), and (3) they scored normally on a test of pitch perception. The remaining 33 participants were amateur musicians that had been playing music several times a week (5 minimum to 10 hours maximum per week was our range of inclusion), for a time duration ranging from one to 26 years at the time of the study. We chose to select young adult participants in order to exclude possible effects of age on grey matter and to focus mainly on the effect of the duration of musical practice. Thus, in order to reveal the dynamic of grey matter changes, the musicians were divided into three groups according to the musical education phases and levels of trainings such as they are cut out in the musical academies in France (Table 1). Thus, the 11 musicians with 1–8 years' musical practice constituted the novice group. These novice musicians were in the process of acquiring basic musical skills (rhythm and music reading), which takes 8 years to be completed. The 11 musicians with 9–14 years' musical practice constituted the intermediate group. In this group, musicians were working on their musical training in preparation for the French final musical diploma. Finally, the 11 musicians with 15 or more years' musical practice constituted the expert musician group who had obtained their French final musical diploma (*'Certificat de fin d'études musicales'*). Thus, within each group, musical proficiencies of our participants are pretty homogeneous. Moreover, we intentionally set up our groups of musicians by choosing various types of instrumental practice (violin, cello, guitar, flute, recorder, trumpet, clarinet and piano) in order to avoid the possible bias induced by a particular type of instrumental practice (in reference to the work of Sluming et al., 2002). Most of our musicians played more than one instrument. The distribution of the primary instrument played by the musicians of our three groups is the following one: 9 strings and 2 winds in novice musician group; 5 strings, 5 winds and 1 pianist in intermediate musician group; 1 string, 5 winds and 5 pianists in expert musician group.

### 2.2. MRI data acquisition

Each participant underwent an MRI examination at the CYCERON center (Caen, France) using the Philips (Eindhoven, The Netherlands) Achieva 3.0T scanner. T1-weighted structural images were acquired using a 3D fast field-echo sequence (3D-T1-FFE sagittal; TR = 20 ms; TE = 4.6 ms; flip angle = 20°; 170 slices; slice thickness = 1 mm; no gap; FOV = 256 × 256 $mm^2$; matrix = 256 × 256; in-plane resolution = 1 × 1 $mm^2$; acquisition time = 9.7 min).

### 2.3. Data preprocessing and statistical analysis

#### 2.3.1. Demographic statistical analyses

One-way ANOVA were run on the demographic data: age and level of education. Mean level of education differed significantly between the nonmusician (16.9 ± 2.17 years) and novice musician groups (14.64 ± 2.80 years), $p < .05$; see Table 1). Mean age differed significantly between the nonmusician (25 ± 3.41 years) and intermediate musician groups (21.09 ± .94 years), and between the novice (24.90 ± 3.80 years) and intermediate musician groups (21.09 ± .94 years, $p < .05$; see Table 1). Consequently, in all the statistical analyses, age and educational level were included as confounding variables.

We performed a Kruskal-Wallis one-way ANOVA in order to test gender distribution and no difference was observed between the four groups.

#### 2.3.2. Anatomical data preprocessing

Imaging data preprocessing and analysis were performed using SPM12 software (Wellcome Trust Center for Neuroimaging, London, UK) implemented in Matlab 7.4. Briefly, individual MRI data were spatially normalized to the Montreal Neurological Institute (MNI) template and segmented to isolate the grey matter partitions

**Table 1**
Mean (±standard deviation) demographic data for each group. For the statistical results of the post hoc 2-by-2 comparisons (Tukey's HSD) only significant differences at $p < .05$ are shown: [a]compared with nonmusician group, [b]compared with novice musician group.

| | Nonmusicians | Novice musicians | Intermediate musicians | Expert musicians |
|---|---|---|---|---|
| Female/male | 5F/6M | 4F/7M | 6F/5M | 3F/8M |
| Age in years (±*SD* and range) | 25 ± 3.41 (21–32) | 24.90 ± 3.80 (21–32) | 21.09 ± 0.94 (20–23)[a,b] | 24.18 ± 3.54 (21–32) |
| Education in years (±*SD* and range) | 16.9 ± 2.17 (14–20) | 14.64 ± 2.80 (10–19)[a] | 14.82 ± 0.60 (14–16) | 15.45 ± 1.13 (14–17) |
| Range of musical practice (mean ± *SD*) | 0 | 1–8 (4.6 yrs ± 2.83) | 9–14 (13.5 yrs ± 0.52) | 15–26 (17.5 yrs ± 3.47) |
| Age of onset of training in years (±*SD* and range) | – | 20.27 ± 5.33 (14–31) | 7.45 ± 0.82 (6–9) | 6.63 ± 1.36 (5–9) |

using the New Segment procedure in SPM12 and the DARTEL toolbox (Ashburner, 2007). Finally, the resulting modulated images (i.e. grey matter (GM) volume), which allow detection of relatively subtle changes in tissue contrasts were smoothed with an 8-mm FWHM isotropic Gaussian kernel. The resulting preprocessed images were masked so as to include only voxels considered as GM in the statistical analyses.

We also obtained the individual total volume of GM, white matter (WM) and cerebrospinal fluid (CSF) that we used to calculated the individual total intracranial volume (TIV) by summing the volume of the three compartments. In the imaging analyses described below, the TIV was used as a covariate to correct for brain volume difference.

*2.3.3. Imaging statistical analyses*

*2.3.3.1. Brain regions modified by musical practice.* First, we performed a regression analysis for the whole sample (44 participants) between the whole-brain GM volumes and the duration of musical practice, controlling for age, educational level and TIV, in order to first, locate the areas where grey matter volumes increased with the duration of musical practice and second, highlight regions where grey matter volumes decreased with the duration of musical practice.

*2.3.3.2. Dynamics of brain regions modified by years of musical training.* Second, to highlight the possible dynamics of grey matter volume modifications related to duration of musical practice, we divided the participants into four groups (Table 1) according to the duration of their musical training and extracted (using a fMRI_ROI toolbox developed in our lab) the grey matter volumes of the areas that were significant (at $p < .001$ uncorrected for multiple comparisons and cluster size $k > 110$ in the regression analysis. Using Statistica software we performed one-way ANOVAs (controlling for age, educational level and TIV) for each region and conduct post hoc comparisons (Tukey's HSD) of grey matter volumes between the four groups (nonmusicians, and novice, intermediate and expert musicians). For the statistical results of the post hoc 2-by-2 comparisons (Tukey's HSD) we discussed significant differences at $p < 0.05$.

*2.3.3.3. Influence of age of onset of musical training.* Finally, to highlight brain areas modified by early and late musical training (before and after age 7 of onset), we performed as a complementary analysis a two-sample *t*-test (at $p < .001$ uncorrected for multiple comparisons and cluster size $k > 110$) on GM images of the 15 musicians who began music before age 7 compared to 18 musicians who began music later (after age 7). To control for duration of musical practice, number of years of practice was included as covariate, and we also set age, educational level and TIV as covariates.

## 3. Results

### 3.1. Brain regions modified by musical practice

Positive correlations between GM volume and the duration of musical practice were found in the left hippocampus, posterior cingulate gyrus and superior temporal cortex, and in the right insular, middle and superior frontal cortices, and supplementary motor area Fig. 1 and Table 1). These same areas were found after controlling for gender, by including this parameter as a nuisance variable in the model (data not shown). No brain area exhibited a negative correlation between GM volume and the duration of musical practice.

### 3.2. Dynamics of brain regions modified by years of musical training

Except for the right insula ($F(3,37) = 2.25$, $p = 0.09$), all these areas present a significant main effect of musical expertise, after controlling for age, educational level and TIV: the left hippocampus ($F(3,37) = 5.25$, $p = 0.003$), posterior cingulate ($F(3,37) = 3.56$, $p = 0.023$) and superior temporal cortex ($F(3,37) = 3.71$, $p = 0.019$), and the right middle and superior frontal cortices ($F(3,37) = 6.39$, $p = 0.001$), and supplementary motor area ($F(3,37) = 5.78$, $p = 0.002$). Post hoc between-group comparisons conducted on the GM volume of the brain regions found in the previous analysis suggested distinct patterns of training-induced structural changes (Fig. 1. Statistical results of the post hoc 2-by-2 comparisons (Tukey's HSD) revealed, for the left hippocampus: significant differences between the nonmusicians and novice ($p = 0.024$), intermediate ($p = 0.0007$) and expert musician ($p = 0.0001$) groups and between the novice musicians and expert musicians ($p = 0.015$); for the left posterior cingulate cortex: significant differences between the nonmusicians and the expert musicians ($p = 0.0002$) and between the intermediate musicians and the expert musicians ($p = 0.036$); for the left superior temporal gyrus: significant differences between nonmusicians and the expert musicians ($p = 0.003$) and between the intermediate musicians and the expert musicians ($p = 0.032$); for the right middle and superior frontal cortices: significant differences between the nonmusicians and intermediate ($p = 0.027$) and expert musician ($p = 0.0001$) groups and between the expert musicians and the novice ($p = 0.009$) and intermediate musicians ($p = 0.036$); and for the supplementary motor area the only significant difference was observed between the nonmusicians and the expert musicians ($p = 0.04$) (Fig. 1) (see Table 2).

### 3.3. Influence of age of onset of musical training

The two-sample *t*-test performed on 15 musicians who began music before age 7 compared to 18 musicians who started music later (after age 7), revealed no significant GM difference, and neither did the reverse comparison. Thus, controlling for the duration of musical practice, no brain area seemed to be specifically influenced by the age of onset of musical training.

## 4. Discussion

Consistent with the evidence that musical training can bring about neural plasticity in the brain (for reviews, see Stewart, 2008; Wan & Schlaug, 2010), the results of the present study showed only significant positive correlations between the duration of musical practice and changes in grey matter volumes in specific

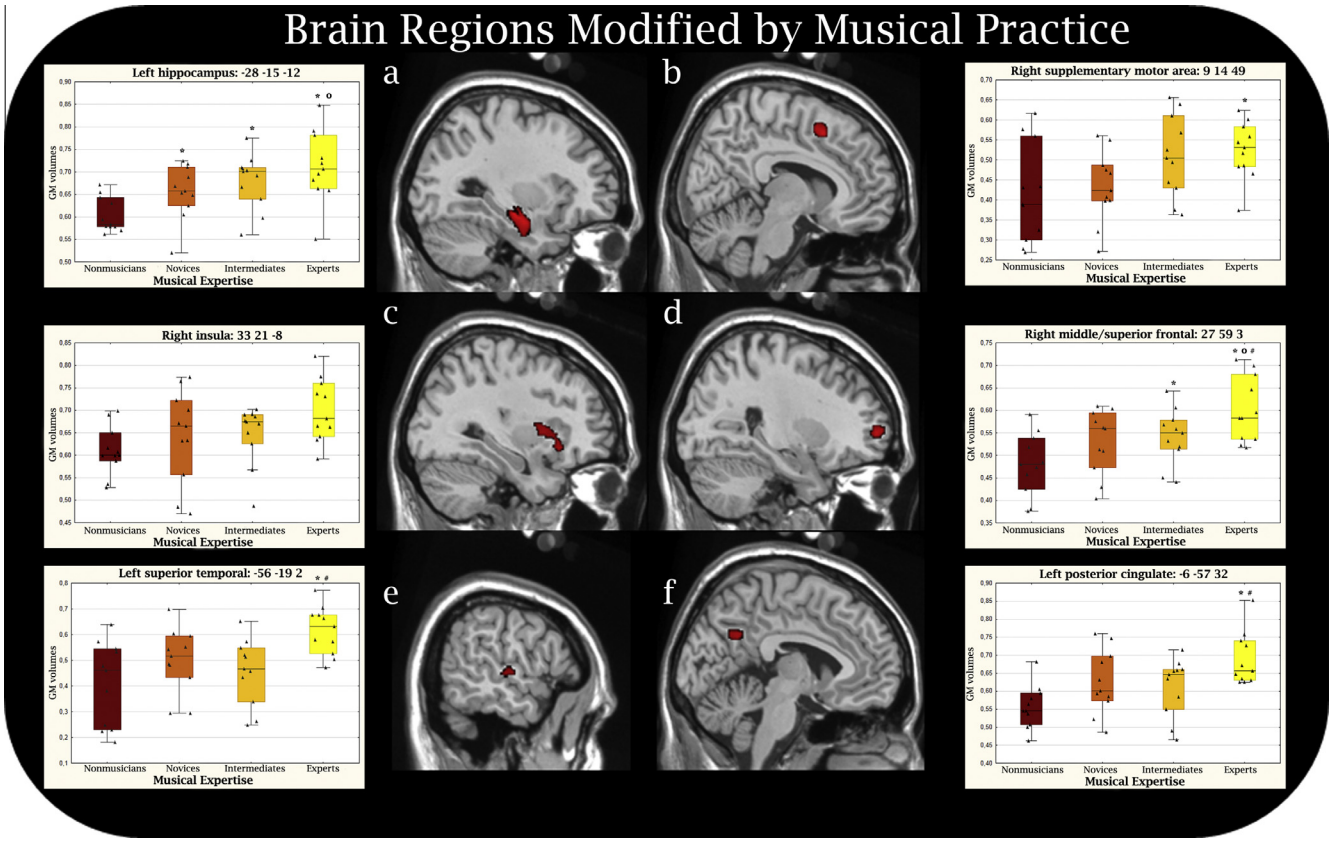


**Fig. 1.** Brain areas in which grey matter density changed with musical expertise. Results displayed the brain areas with significant positive correlation at $p < .001$ uncorrected for multiple comparisons and cluster size $k > 110$ between the duration of musical practice and GM volumes (controlling for age, educational level and TIV) and boxplot representations of GM volumes of (a) the left hippocampus, (b) right supplementary area, (c) right insula, (d) right superior and middle frontal cortex, (e) left superior temporal cortex and (f) left posterior cingulate cortex for each group. Box plots represent data of each group with quartiles (upper values 75%, median, and lower values 25%). The whiskers show range of the data and blobs raw values. For statistical results of the post hoc 2-by-2 comparisons (Tukey's HSD) only significant differences are shown: * compared with the nonmusician group, o compared with the novice musician group and # compared with the intermediate musician group.

**Table 2**
Brain areas in which GM volume changed with musical expertise. Location and MNI *x*, *y*, *z* coordinates (in mm) of peaks of significant grey matter density increases revealed by the analysis of variance with age, years of education and TIV as covariates. In all the areas listed, changes in grey matter volume are statistically significant at the $p < .001$ uncorrected level for multiple comparisons and cluster size $k > 110$.

| Anatomical location | Cluster size (in voxels) | *x* | *y* | *z* | *Z* score |
|---|---|---|---|---|---|
| Left hippocampus | 434 | −28 | −15 | −12 | 4.77 |
| Right supplementary motor area | 114 | 9 | 14 | 49 | 4.43 |
| Right superior and middle frontal cortex | 211 | 27 | 59 | 3 | 4.22 |
| Left posterior cingulate cortex | 165 | −6 | −57 | 32 | 4.02 |
| Right insula | 177 | 33 | 21 | −8 | 3.70 |
| Left superior temporal | 122 | −56 | −19 | 2 | 3.47 |

brain regions that have already been found to be sensitive to musical training or expertise (Bermudez et al., 2009; Groussard et al., 2010a; Hutchinson, Lee, Gaab, & Schlaug, 2003; Hyde et al., 2009; James et al., 2014; Schneider et al., 2002). The comparison of subjects with different levels of musical expertise enabled us to extend our understanding of brain plasticity in musicians, by allowing us to analyze different grey matter changes within each brain region in the course of musical training.

### 4.1. Different structural changes across musical practice

Our analyses indicated that the grey matter modifications associated with musical practice differ according to brain regions (Fig. 1). These changes appear to be already noticeable in novice musicians and were greater in the intermediate and expert musician group in the right middle and superior frontal regions and left hippocampus, whereas structural modifications in the left superior temporal, posterior cingulate and right supplementary motor areas seem to be less linear and appear only after several years of musical practice. Given that we controlled for age and education level, these results cannot be ascribed to normal cerebral maturation. So, these findings suggest that experience-dependent shifts within cognitive processes take place during musical training (Bermudez et al., 2009; Hyde et al., 2009).

#### 4.1.1. Left superior temporal gyrus

The left superior temporal gyrus is the structure where grey matter density greater the most (50%) in expert musicians versus nonmusicians. Previous studies observed an increase of GM in the left superior temporal lobes in musicians (Gaser & Schlaug,

2003; Schneider et al., 2002; James et al., 2014) compared to nonmusicians suggesting a direct effect of musical practice on this area. Moreover, functional imaging studies have revealed that both superior temporal areas were involved in the processing of melody (Bengtsson & Ullén, 2006) and the left one in musical semantic memory (Groussard et al., 2010b). We thus can hypothesize that throughout their training, musicians gradually develop the ability to decode the perceptual features of tunes (melody) and to memorize these features, in order to obtain a unique representation of each musical encounter. Moreover, previous research has suggested that the temporal structure is central to the recognition of familiar melodies. This brain area is thought to sustain the musical lexicon and could be involved in perceptual musical memories (Peretz et al., 2009). Given that the musical lexicon is gradually enriched through training and through listening to tunes, our results suggest that the grey matter differences in the left temporal area can be attributed to musical expertise.

#### 4.1.2. Left posterior cingulate cortex

We observed a 23.5% enlargement in grey matter volume in the left posterior cingulate cortex of the expert musicians, compared with the nonmusicians. Hyde et al. (2009) previously observed a grey matter increase in the posterior cingulate cortex in children after 15 month of musical training and interpreted this fact as a consequence of integration of sensory information (visual) and emotional content occurring during the learning to read musical notation. Moreover, in musical memory studies, this area appears to be activated during familiarity tasks featuring well-known songs (Groussard et al., 2010b; Satoh, Takeda, Nagata, Shimosegawa, & Kuzuhara, 2006) and could underlie autobiographical memories associated with musical excerpts (Ford, Addis, & Giovanello, 2011).

#### 4.1.3. Right insular cortex

We observed an enlargement of grey matter in the insula cortex with the duration of musical practice but no clear pattern of modification appears regarding the level of musical expertise, nevertheless we could observe 14.6% difference between the expert musicians and the nonmusicians. This area is thought to reflect the emotional aspects of music processing (Koelsch, 2010; Koelsch, Fritz, Schulze, Alsop, & Schlaug, 2005). A positive correlation has been observed between the activation of the insula and the intensity of the thrills induced in musicians by favorite pieces of classical music (Blood & Zatorre, 2001). We can assume that, musicians hone not just their technical prowess but also their emotional sensitivity to music. In the course of their musical experience, musicians develop the ability (their sensitivity) to perceive the emotional content inherent to different pieces of music and, in turn, the ability to communicate that emotion to their audience.

#### 4.1.4. Right supplementary motor area

An increase in grey matter of the right supplementary motor area only seems to appear after 15 or more years of musical practice, amounting to 26% between the expert musicians and the nonmusicians. This structure, together with the premotor cortex, had previously shown a greater GM volume in musicians vs. nonmusicians (Gaser & Schlaug, 2003). This cerebral area was also observed in musical studies focusing on the processing of sequential temporal structures in passive rhythm perception (Bengtsson et al., 2009). fMRI studies, revealed that in pianists the supplementary motor area could be involved in both pitch and timing repetition during listening and performance tasks (Brown et al., 2013), as well as in the rhythmic and melodic musical improvisation (Manzano & Ullén, 2012) confirming its probable implication during musical training. On the other hand, some studies suggested that when musicians reach a professional status, the activity of motor structures decrease, which is likely to reflect motor automatization or specialization (Gaser & Schlaug, 2003; James et al., 2014).

#### 4.1.5. Right superior and middle frontal gyri

Our study shows that, in the right superior and middle frontal gyri, grey matter volumes seems to undergo a gradual increase, to reach a 25% difference between the expert musicians and the nonmusicians. These structures have already been found in musical studies, and appear engaged in musical episodic retrieval (Platel, Baron, Desgranges, Bernard, & Eustache 2003) and also recruited by self-referential processes associated with music (Zatorre, Halpern, & Bouffard, 2010), that are high-level cognitive processes developed with musical experience. One recent study explored brain structures that are activated when musicians play in an ensemble, and suggested that the cognitive empathy needed to take account of the other musicians and to play a piece of music together accurately is largely mediated by frontal areas BA 10/11 (Babiloni et al., 2012). Thus, we can hypothesize that the gradual modification in the right superior and middle frontal cortex with the musical expertise appears because playing in a music ensemble requires mastery of one's instrument synchronized with the other players, which is acquired progressively with the years of practice.

#### 4.1.6. Left anterior hippocampus

We observed a gradual increase between duration of musical practice and grey matter volumes in the anterior part of the left hippocampus, suggesting a specific impact of musical practice on this area. This difference reached 17.8% with a greater grey matter volume for the expert musicians, compared with the nonmusicians.

Hippocampal grey matter changes were already noticeable in the novice vs. nonmusicians. This structure is closely involved in context-dependent episodic or autobiographical memory (Burgess, Maguire, & O'Keefe, 2002; Viard et al., 2007), suggesting that musicians construct specific memories relating to their musical experiences (e.g., a particular melody played during a specific concert). Thus, musicians' memories of music may be more detailed, vivid and emotional than those of nonmusicians (Groussard et al., 2010a). Using grey matter volume analysis, Rajah, Kromas, Han, and Pruessner (2010) observed the involvement of the anterior part of the hippocampus in binding spatial and temporal contextual details with item information during encoding and/or retrieval - an ability that is also required in musical practice (e.g., reading a music score). Given the high intensity of auditory-visual encoding processes that characterizes music learning, our findings fit well with this assumption.

Moreover, training-related neurogenesis has been found in the hippocampi of both animals and humans (Eriksson et al., 1998; Fotuhi, Do, & Jack 2012; Kempermann, Gast, & Gage, 2002 for review). In humans, modifications in hippocampal grey matter have been observed such as in London taxi drivers (Maguire et al., 2000) and medical students taking examinations (Draganski et al., 2006), as well as in older people following a period of intense learning (Boyke et al., 2008), suggesting the *biological engraving* of new learning. These observations, taken together with the present results, suggest that musical practice could influence neurogenesis in the left hippocampus, whatever the level of musical expertise or the age of onset, and could possibly reflect an increase in memory faculties.

## 5. Profits and limits

We carried out a global regression analysis in order to give an account of the effect of years of musical practice on GM volume. Given that we controlled for age and education level, and that all

participants were young adults, our results cannot be ascribed to normal cerebral maturation or aging effect. In addition, the major interest of including a group of strictly nonmusicians in our experimental sample is precisely to be able to reveal early GM modifications related to the musical practice. Then, the ANOVA carried out makes it possible to show the contribution of each group in the GM modifications, and allows us to analyze grey matter changes within each brain region in relation with the number of years of musical training. One can however wonder whether the results of the regression analysis could mainly reflect the contribution of the nonmusicians group. We thus carried out the regression analysis (data not shown) on the 3 groups of musicians only, with TIV, Age, and Education level as covariates, and we obtained most of the cerebral areas already found in our previous analysis (that included nonmusicians). Unlike James et al. (2014) we did not observe any decrease in sensorimotor areas, possibly because our study included amateur musicians. All the musician in our study (from "novice' to "expert") practiced 5–10 h per week (selection criteria) whereas in the study of James et al. (2014) musicians who were professional pianists practiced more than 10 h. It is thus possible that the decrease in sensorimotor areas GM density needs a more intensive musical practice than that of non-professional musicians.

While most studies included only pianists, we deliberately decided not to select our subjects on instrumental criteria (in reference to the work of Sluming et al., 2002) so as to limit the influence of a particular type of instrumental practice. The global repartition of instrumental practice is quite homogeneous in our sample of musicians (see 2.1 Participants), and most of our musician subjects, after learning a primary instrument had finally become multi-instrumentists. Our results thus reflect the cerebral effects of musical rather than instrumental training. However, it would be definitely interesting to have a larger sample of different kind of instrumentalists in order to learn whether particular instrumental practices differently modify GM volume during time.

Concerning the possible effect of gender repartition in our results, we carried out a Kruskall-Wallis ANOVA, which does not show any significant difference between male and female repartition in our experimental group. We also checked that the results of the main correlation between musical practice and GM volume were not modified by including gender as covariate in our analysis (data not shown). However, we did a comparison (two-sample) between men and women participants by putting in covariate the number of years of practice. We found a significant greater GM volume in men in motor areas: the cerebellum and the paracentral lobule (data not shown), while the opposite comparison did not show any such significant difference in women vs. men. As some authors have already observed (Hutchinson et al., 2003), if we compare men and women populations with the duration of musical practice as a covariate, grey matter differences are found only in men in motor area: cerebellum and paracentral lobule. The reason of such effect is unknown and needed future studies.

Clearly, the cross-sectional design of this study limits the interpretations regarding the dynamics of grey matter modifications with duration of musical practice. These could be only confirmed with longitudinal studies carried out in nonmusicians who would learn a musical instrument over a long period of time (e.g. 15 or more years of musical practice), which is quite a challenging task in a research context. Nevertheless, our results do provide substantial evidence that localized, structural modifications and adaptations occur in response to the long-term musical training and acquisition of skills specifically needed to play an instrument. As the two-sample t–test performed on 15 musicians who began music before age 7 compared to 18 musicians who started music later (after age 7), revealed no significant GM difference (as the reverse comparison), our results are likely to be explained by the number of years of musical practice rather than to reflect the influence of the onset of musical education taking place at a "sensitive" period of life (White, Hutka, Wiliams, & Moreno, 2013). In other word, although some authors observed an effect of early-training music (for example, Bailey et al., 2014) our data suggested that even if musical training is begin after age of 7, musical practice could also modified structural brain. Finally, it is difficult to isolate from the structural brain modifications observed, the respective impact of the age of onset and duration of musical practice, but these two variables have undoubtedly an influence and are probably additional.

## 6. Conclusion

The present findings illustrate the dynamics of structural brain changes related to musical practice. While neural plasticity occurs in some regions (left hippocampus and right middle and superior frontal), as soon as an individual engages in music learning, grey matter modifications in other brain areas (left posterior cingulated cortex, superior temporal areas and right supplementary motor area and insula cortex) require several years of practice. These differential dynamics of structural neuroplasticity according to brain regions can be attributed to two nonexclusive mechanisms. First, these brain regions may differ in terms of their intrinsic plastic properties (e.g., the hippocampus and its mechanism of neurogenesis). Second, the differential dynamics of change may reflect the involvement of different cognitive functions at different stages in music learning: initial improvements in motor, visual and perceptual skills and the progressive enhancement of the higher-level cognitive processes, such as semantic or episodic memory, meta-representation (emotional interpretation and musical expressiveness) and executive functions that are essential for playing in a music ensemble.

Further investigations are now required to determine whether, from a more clinical perspective, musical practice could increase the so-called cognitive reserve in healthy aging and, perhaps, delay the emergence of cognitive decline in older people (Verghese et al., 2003).

## Acknowledgments

We would like to thank all the members of the INSERM U1077 team, the staff at the Cyceron center for their help with data acquisition, Anne-Lise Pitel for providing such helpful comments and Elizabeth Portier for reviewing the English style. This study was supported by a grant from the Caisse d'Epargne Normandie foundation.

## Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bandc.2014.06.013.

## References

Altenmüller, E. (2008). Neurology of musical performance. *Clinical Medicine, 8*, 410–413.

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *Neuroimage, 38*(1), 95–113.

Babiloni, C., Buffo, P., Vecchio, F., Marzano, N., Del Percio, C., Spada, D., et al. (2012). Brains "in concert": Frontal oscillatory alpha rhythms and empathy in professional musicians. *Neuroimage, 60*, 105–116.

Bailey, J. A., Zatorre, R., & Penhune, V. B. (2014). Early musical training is linked to gray matter structure in the ventral premotor cortex and auditory-motor rhythm synchronization performance. *Journal of Cognitive Neuroscience, 26*, 755–767.

Bengtsson, S. L., & Ullén, F. (2006). Dissociation between melodic and rhythmic processing during piano performance from musical scores. *Neuroimage, 30*, 272–284.

Bengtsson, S. L., Ullén, F., Ehrsson, H. H., Hashimoto, T., Kito, T., Naito, E., et al. (2009). Listening to rhythms activates motor and premotor cortices. *Cortex, 45*(1), 62–71. http://dx.doi.org/10.1016/j.cortex.2008.07.002.

Bermudez, P., Lerch, J. P., Evans, A. C., & Zatorre, R. J. (2009). Neuroanatomical correlates of musicianship as revealed by cortical thickness and voxel-based morphometry. *Cerebral Cortex, 19*, 1583–1596.

Blood, A. J., & Zatorre, R. J. (2001). Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *PNAS, 98*, 11818–11823.

Boyke, J., Driemeyer, J., Gaser, C., Büchel, C., & May, A. (2008). Training-induced brain structure changes in the elderly. *Journal of Neuroscience, 28*, 7031–7035.

Brown, R. M., Chen, J. L., Hollinger, A., Penhune, V. B., Palmer, C., & Zatorre, R. J. (2013). Repetition suppression in auditory-motor regions to pitch and temporal structure in music. *Journal of Cognitive Neuroscience, 25*(2), 313–328. http://dx.doi.org/10.1162/jocn_a_00322.

Burgess, N., Maguire, E. A., & O'Keefe, J. (2002). The human hippocampus and spatial and episodic memory. *Neuron, 35*, 625–641.

Draganski, B., Gaser, C., Busch, V., Schuierer, G., Bogdahn, U., & May, A. (2004). Neuroplasticity: Changes in grey matter induced by training. *Nature, 427*, 311–312.

Draganski, B., Gaser, C., Kempermann, G., Kuhn, H. G., Winkler, J., Büchel, C., et al. (2006). Temporal and spatial dynamics of brain structure changes during extensive learning. *Journal of Neuroscience, 26*, 6314–6317.

Draganski, B., & May, A. (2008). Training-induced structural changes in the adult human brain. *Behavioural Brain Research, 192*, 137–142.

Engvig, A., Fjell, A. M., Westlye, L. T., Moberget, T., Sundseth, Ø., Larsen, V. A., et al. (2010). Effects of memory training on cortical thickness in the elderly. *Neuroimage, 52*, 1667–1676.

Eriksson, P. S., Perfilieva, E., Björk-Eriksson, T., Alborn, A. M., Nordborg, C., Peterson, D. A., et al. (1998). Neurogenesis in the adult human hippocampus. *Nature Medicine, 4*, 1313–1317.

Fauvel, B., Groussard, M., Eustache, F., Desgranges, B., & Platel, H. (2013). Neural implementation of musical expertise and cognitive transfers: Could they be promising in the framework of normal cognitive aging? *Frontiers in Human Neuroscience, 7*, 693. http://dx.doi.org/10.3389/fnhum.2013.00693.

Fauvel, B., Groussard, M., Chételat, G., Fouquet, M., Landeau, B., Eustache, F., et al. (2014). Morphological brain plasticity induced by musical expertise is accompanied by modulation of functional connectivity at rest. *Neuroimage, 90*, 179–188. http://dx.doi.org/10.1016/j.neuroimage.2013.12.065.

Ford, J. H., Addis, D. R., & Giovanello, K. S. (2011). Differential neural activity during search of specific and general autobiographical memories elicited by musical cues. *Neuropsychologia, 49*, 2514–2526.

Fotuhi, M., Do, D., & Jack, C. (2012). Modifiable factors that alter the size of hippocampus with ageing. *Nature Reviews Neurology, 8*, 189–202.

Gaser, C., & Schlaug, G. (2003). Brain structures differ between musicians and non-musicians. *Journal of Neuroscience, 23*, 9240–9245.

Groussard, M., La Joie, R., Rauchs, G., Landeau, B., Chételat, G., Viader, F., et al. (2010a). When music and long-term memory interact: Effects of musical expertise on functional and structural plasticity in the hippocampus. *PLoS ONE, 5*, e13225.

Groussard, M., Rauchs, G., Landeau, B., Viader, F., Desgranges, B., Eustache, F., et al. (2010b). The neural substrates of musical memory revealed by fMRI and two semantic tasks. *Neuroimage, 53*, 1301–1309.

Hutchinson, S., Lee, L. H.-L., Gaab, N., & Schlaug, G. (2003). Cerebellar volume of musicians. *Cerebral Cortex, 13*, 943–949.

Hyde, K. L., Lerch, J., Norton, A., Forgeard, M., Winner, E., Evans, A. C., et al. (2009). Musical training shapes structural brain development. *Journal of Neuroscience, 29*, 3019–3025.

James, C. E., Oechslin, M. S., Van De Ville, D., Hauert, C.-A., Descloux, C., & Lazyras, F. (2014). Musical training yields opposite effects on grey matter density in cognitive versus sensorimotor networks. *Brain Structure and Function, 219*, 353–366. doi:10.1007/s00429-013-0504-z.

Jancke, L. (2009). The plastic human brain. *Restorative Neurology and Neuroscience, 27*, 521–538.

Kempermann, G., Gast, D., & Gage, F. H. (2002). Neuroplasticity in old age: Sustained fivefold induction of hippocampal neurogenesis by long-term environmental enrichment. *Annals of Neurology, 52*, 135–143.

Koelsch, S. (2010). Towards a neural basis of music-evoked emotions. *Trends in Cognitive Sciences, 14*, 131–137.

Koelsch, S., Fritz, T., Schulze, K., Alsop, D., & Schlaug, G. (2005). Adults and children processing music: An fMRI study. *Neuroimage, 25*, 1068–1076.

Luders, E., Gaser, C., Jancke, L., & Schlaug, G. (2004). A voxel-based approach to gray matter asymmetries. *Neuroimage, 22*, 656–664.

Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S., et al. (2000). Navigation-related structural change in the hippocampi of taxi drivers. *PNAS, 97*, 4398–4403.

Manzano, O., & Ullén, F. (2012). Activation and connectivity patterns of presupplementary and dorsal premotor areas during free improvisation of melodies and rythms. *NeuroImage, 63*, 272–280.

May, A. (2010). Experience-dependent structural plasticity in the adult human brain. *Trends in Cognitive Sciences, 15*, 475–482.

Oechslin, M. S., Van De Ville, D., Lazeyras, F., & James, J. E. (2012). Degree of musical expertise modulates higher order brain functioning. *Cerebral Cortex*. http://dx.doi.org/10.1093/cercor/bhs206.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9*, 97–113.

Park, I. S., Lee, K. J., Han, J. W., Lee, N. J., Lee, W. T., Park, K. A., et al. (2009). Experience-dependent plasticity of cerebellar vermis in basketball players. *Cerebellum, 8*, 334–339.

Peretz, I., Gosselin, N., Belin, P., Zatorre, R. J., Plailly, J., & Tillmann, B. (2009). Music lexical networks: The cortical organization of music recognition. *NYAS, 1169*, 256–265.

Platel, H., Baron, J.-C., Desgranges, B., Bernard, F., & Eustache, F. (2003). Semantic and episodic memory of music are subserved by distinct neural networks. *Neuroimage, 20*, 244–256.

Rajah, M. N., Kromas, M., Han, J. E., & Pruessner, J. C. (2010). Group differences in anterior hippocampal volume and in the retrieval of spatial and temporal context memory in healthy young versus older adults. *Neuropsychologia, 48*, 4020–4030.

Satoh, M., Takeda, K., Nagata, K., Shimosegawa, E., & Kuzuhara, S. (2006). Positron-emission tomography of brain regions activated by recognition of familiar music. *American Journal of Neuroradiology, 27*, 101–106.

Schneider, P., Scherg, M., Dosch, H. G., Specht, H. J., Gutschalk, A., & Rupp, A. (2002). Morphology of Heschl's gyrus reflects enhanced activation in the auditory cortex of musicians. *Nature Neuroscience, 5*, 688–694.

Schulze, K., Mueller, K., & Koelsch, S. (2011). Neural correlates of strategy use during auditory working memory in musicians and non-musicians. *European Journal of Neuroscience, 33*, 189–196.

Seung, Y., Kyong, J.-S., Woo, S.-H., Lee, B.-T., & Lee, K.-M. (2005). Brain activation during music listening in individuals with or without prior music training. *Neuroscience Research, 52*, 323–329.

Sluming, V., Barrick, T., Howard, M., Cezayirli, E., Mayes, A., & Roberts, N. (2002). Voxel-based morphometry reveals increased gray matter density in Broca's area in male symphony orchestra musicians. *Neuroimage, 17*, 1613–1622.

Steele, C. J., Bailey, J. A., Zatorre, R. J., & Penhune, V. B. (2013). Early musical training and white-matter plasticity in the corpus callosum: Evidence for a sensitive period. *Journal of Neuroscience, 33*, 1282–1290.

Stewart, L. (2008). Do musicians have different brains? *Clinical Medicine, 8*, 304–308.

Taubert, M., Draganski, B., Anwander, A., Müller, K., Horstmann, A., Villringer, A., et al. (2010). Dynamic properties of human brain structure: Learning-related changes in cortical areas and associated fiber connections. *Journal of Neuroscience, 30*, 11670–11677.

Verghese, J., Lipton, R. B., Katz, M. J., Hall, C. B., Derby, C. A., Kuslansky, G., et al. (2003). Leisure activities and the risk of dementia in the elderly. *New England Journal of Medicine, 348*, 2508–2516.

Viard, A., Piolino, P., Desgranges, B., Chételat, G., Lebreton, K., Landeau, B., et al. (2007). Hippocampal activation for autobiographical memories over the entire lifetime in healthy aged subjects: An fMRI study. *Cerebral Cortex, 17*, 2453–2467.

Wan, C. Y., & Schlaug, G. (2010). Music making as a tool for promoting brain plasticity across the life span. *Neuroscientist, 16*, 566–577.

White, E. J., Hutka, S. A., Wiliams, L. J., & Moreno, S. (2013). Learning, neural plasticity and sensitive periods: Implications for language acquisition, music training and transfer across the lifespan. *Frontiers in Systems Neuroscience, 20*(7), 90. http://dx.doi.org/10.3389/fnsys.2013.00090.

Zatorre, R. J., Fields, R. D., & Johansen-Berg, H. (2012). Plasticity in gray and white: Neuroimaging changes in brain structure during learning. *Nature Neuroscience, 15*, 528–536.

Zatorre, R. J., Halpern, A. R., & Bouffard, M. (2010). Mental reversal of imagined melodies: A role for the posterior parietal cortex. *Journal of Cognitive Neuroscience, 22*, 775–789.

Behavioral/Systems/Cognitive

# Musical Training Induces Functional Plasticity in Human Hippocampus

**Marcus Herdener,[1] Fabrizio Esposito,[2,3] Francesco di Salle,[3] Christian Boller,[4] Caroline C. Hilti,[1] Benedikt Habermeyer,[6] Klaus Scheffler,[4] Stephan Wetzel,[5] Erich Seifritz,[1,7,8] and Katja Cattapan-Ludewig[1,7]**

[1]University Hospital of Psychiatry, University of Bern, 3000 Bern, Switzerland, [2]Department of Neuroscience, University of Naples "Federico II," 80131 Naples, Italy, [3]Department of Cognitive Neuroscience, University of Maastricht, 6200 MD Maastricht, The Netherlands, [4]Division of Radiological Physics, Institute of Radiology, University of Basel Hospital, and [5]Department of Medical Radiology, Institute of Neuroradiology, University of Basel Hospital, 4031 Basel, Switzerland, [6]Department of Psychiatry, University of Basel, 4025 Basel, Switzerland, [7]Sanatorium Kilchberg, 8802 Kilchberg, Switzerland, and [8]Psychiatric University Hospital, University of Zurich, 8032 Zurich, Switzerland

**Training can change the functional and structural organization of the brain, and animal models demonstrate that the hippocampus formation is particularly susceptible to training-related neuroplasticity. In humans, however, direct evidence for functional plasticity of the adult hippocampus induced by training is still missing. Here, we used musicians' brains as a model to test for plastic capabilities of the adult human hippocampus. By using functional magnetic resonance imaging optimized for the investigation of auditory processing, we examined brain responses induced by temporal novelty in otherwise isochronous sound patterns in musicians and musical laypersons, since the hippocampus has been suggested previously to be crucially involved in various forms of novelty detection. In the first cross-sectional experiment, we identified enhanced neural responses to temporal novelty in the anterior left hippocampus of professional musicians, pointing to expertise-related differences in hippocampal processing. In the second experiment, we evaluated neural responses to acoustic temporal novelty in a longitudinal approach to disentangle training-related changes from predispositional factors. For this purpose, we examined an independent sample of music academy students before and after two semesters of intensive aural skills training. After this training period, hippocampal responses to temporal novelty in sounds were enhanced in musical students, and statistical interaction analysis of brain activity changes over time suggests training rather than predisposition effects. Thus, our results provide direct evidence for functional changes of the adult hippocampus in humans related to musical training.**

## Introduction

The ability to make music to a professional standard implies a high degree of performance, which is acquired after years of intensive training and is one of the most complex human achievements involving various brain regions (Peretz, 2006). The musician's brain is thus regarded as a suitable model to study neuroplastic changes (Munte et al., 2002). However, the effect of musical expertise acquired through years of intensive training on functional properties of the hippocampus remained elusive, although animal models show that the hippocampus formation is particularly susceptible to neuroplastic changes modulated by various environmental factors and learning processes (Kempermann et al., 1997; Lledo et al., 2006).

Hippocampal plasticity in humans in general has mainly been inferred indirectly, by measuring the structural changes of this region with volumetric magnetic resonance imaging and relating it to training tasks involving memory functions (Maguire et al., 2000; Draganski et al., 2006). However, evidence for altered functional measures directly reflecting changes in hippocampal processing as induced by environmental factors or training is missing.

In addition to its outstanding role for memory and spatial navigation (Maguire, 2001; Ekstrom et al., 2003), the hippocampus has been suggested to be involved in novelty detection (Knight, 1996; Strange et al., 1999). Hippocampal novelty detection is based on a comparison of actual sensory inputs with stored stimulus patterns (Gray and Rawlins, 1986; Strange and Dolan, 2001; Vinogradova, 2001; Kumaran and Maguire, 2007a).

Music consists of precisely patterned sequences of sounds, and the ability to identify isochronous temporal intervals (and temporal variations) and to synchronize precisely with sensory information is a prerequisite for playing in an ensemble. A fine-tuning of aural skills in professional musicians is achieved by a sophisticated ear training that musical students receive during their academic education and is regarded as crucial component of their vocational formation.

We thus hypothesized that the hippocampus might be critically involved in detection of novelty of temporal structure in the auditory domain and that training of aural skills could modulate

Received Sept. 11, 2009; revised Nov. 17, 2009; accepted Dec. 1, 2009.

This work was supported by the Stiftung für klinische neuropsychiatrische Forschung, Bern, Switzerland. We thank Elke Hofmann (Hochschule für Musik, Basel, Switzerland) and Hans Peter Weber (Schola Cantorum, Basel, Switzerland) for their help in recruiting professional musicians and students from the "Musik-Akademie Basel" for evaluating the aural skills of music students with music dictation and, together with Felipe Cattapan, for helpful discussions.

Correspondence should be addressed to Dr. Marcus Herdener, Max Planck Institute for Biological Cybernetics, Spemannstrasse 41, 72076 Tübingen, Germany. E-mail: marcus.herdener@tuebingen.mpg.de.

DOI:10.1523/JNEUROSCI.4513-09.2010

 0270-6474/10/301377-08$15.00/0

the detection of temporal novelty of acoustic signals in the hippocampus.

We used functional magnetic resonance imaging (fMRI) to test these two hypotheses in two independent experiments. In a first experiment (experiment 1), by presenting an acoustic temporal mismatch paradigm (see Fig. 1) to subjects with different backgrounds of musical training (professional musicians and musical laypersons), we aimed at testing for the involvement of hippocampus in acoustic novelty detection and its putative functional modulation by musical experience using a cross-sectional design. A cross-sectional design, however, leaves open the fundamental question whether observed differences between groups are related to talent or training. Thus, in a second experiment (experiment 2) using a longitudinal design, we examined an independent sample of music academy students before and after academic aural skills training in comparison with students of nonmusical faculties to specify the impact of musical training on hippocampal acoustic novelty detection.

More generally, by investigating auditory novelty detection in musicians' brains, we aimed at studying whether the adult human hippocampus is subject to functional plasticity induced by training.

## Materials and Methods

Both experiments have been approved by the local ethics committee, and all subjects provided written informed consent to participate.

### *Subjects*

In the cross-sectional experiment, we examined seven professional musicians who were professional experts in ear training and seven nonmusicians (for subject characterization, see supplemental Table 1, available at www.jneurosci.org as supplemental material), matched for age and gender. Most of the musicians worked as lecturers at a music academy. In the longitudinal experiment, 19 musical students and 21 control students of other (nonmusical) faculties were examined (supplemental Table 2, available at www.jneurosci.org as supplemental material). The musical students were recruited from the "Musik-Akademie Basel (Hochschule fuer Musik and Schola Cantorum Basiliensis)"; the students of other faculties were recruited at the University of Basel (Basel, Switzerland).

The musical students of this study were at the beginning of their academic music studies at Hochschule fuer Musik, Basel, or at the Schola Cantorum Basiliensis. The students underwent intensive ear training with at least three semester hours lessons per week in ear training (see also supplemental material, available at www.jneurosci.org). The students were also requested by their lecturers to improve their aural skills by training at home. In our study, we tested the musical students two times, at the beginning and at the end of their first two semesters of the ear training. The mean time interval between the two test sessions was 217 d (SD, 16).

Students from other faculties, who, however, did not receive any aural skill training between the two longitudinal evaluation time points in addition to the formal education specific to their individual fields of study, served as a matched control group (mean interval between sessions, 198 d; SD, 20).

### *Auditory stimulation during functional imaging*

Subjects (experiments 1 and 2) were exposed to an acoustic temporal mismatch paradigm [similar to that of Rüsseler et al. (2001)] via MR-compatible headphones (Commander XG; Resonance Technology). In detail, sine tones (50 ms duration; 5 ms linear rise and fall times; carrier frequency, 1 kHz) were adjusted in amplitude according to the subjects' subjective feedback of sounds being clearly distinguishable from scanner noise background while not being experienced as unpleasant [no significant group differences in sound pressure level (SPL); experiment 1: mean SPL, 89 dB; $t$ test (df 12), $p = 0.17$; experiment 2: mean SPL, 83.5 dB; ANOVA, $p = 0.43$]. The stimuli were presented with a standard stimulus onset asynchrony (SOA) of 150 ms. Shorter SOAs with three degrees of deviance (142, 130, and 100 ms) (see Fig. 1) (example audio files are presented in the supplemental material, available at www.jneurosci.org) occurred every 16–20 s, representing the novel deviant temporal structure of acoustic input. In other words, on average, 1 of 120 sine tones represented temporal novelty. Each deviant condition was repeated 12 times in a pseudorandomized order (experimental duration in total, 11 min 26 s). To avoid activation caused by response selection, planning, or working memory, subjects were not required to make any sort of response during the experiment. Moreover, they were instructed to watch a silent movie and to ignore the sounds presented binaurally via headphones. This way, subjects performed an incidental task so as to avoid any explicit judgment on auditory inputs. In our experiments, we thus focused on automatic or task-irrelevant novelty detection in the auditory domain (and its plasticity related to musical training), because automaticity is an important property of an efficient novelty discrimination system allowing the brain to rapidly and effortlessly detect change in the environment (Sokolov, 1963; Brown and Bashir, 2002; Yamaguchi et al., 2004; Kumaran and Maguire, 2007b). Note that this approach is similar to experimental conditions commonly used in electrophysiological studies to test for the neural correlates of preattentive oddball detection in the acoustic domain (for review, see Näätänen and Escera, 2000; Näätänen et al., 2001) and to identify functional differences in automatic auditory processing in musicians and nonmusicians on a cortical (for review, see Munte et al., 2002) and subcortical level (Parbery-Clark et al., 2009).

### *Behavioral testing of musical skills*

To behaviorally assess basic temporal sound processing facilities, subjects of the cross-sectional study performed a standardized test measuring musical abilities after functional imaging, which required the detection of small deviances in short melodies in a forced-choice task [Advanced Measures of Music Audiation (AMMA) test] (Gordon, 1998) (see supplemental Table 1, available at www.jneurosci.org as supplemental material).

For evaluation of aural skills of students in the longitudinal study, they participated in a music dictation, probably the most common measure to evaluate musical aural skills at the university level. At two times (at the beginning at the end of our study period), a piece of music (cut to "takes" of a few seconds) was presented to the students by headphones. The students had to write the musical scores of the soprano and the bass voice. The scores were judged by university teachers at Hochschule fuer Musik by giving credits corresponding to each correct measure in each voice. The achieved and the possible credits of the dictation were converted to a percentage measure of correct answers. The following parts of music have been used for the dictation: (1) W. A. Mozart: symphony KV 319, 2d movement, measures 1–18; (2) J. S. Bach: cantata BWV 119, no. 5 (alto aria), measures 1–13. The degree of difficulty of these two music parts was comparable.

The music dictations took place outside the MR room in the rooms of Hochschule fuer Musik. The students were familiar with the situation of the music dictation. Sixteen (of 19) students participated at the baseline dictation ("Mozart"), and 17 after their two semesters of training (follow-up, "Bach").

### *Data acquisition*

For the cross-sectional experiment (experiment 1), fMRI data were acquired on a 1.5 T standard clinical MRI scanner (Siemens) equipped with an Espree gradient system and a circularly polarized radio frequency headcoil. Data from the longitudinal experiment (experiment 2) were acquired with a 3 T MRI scanner (Siemens) equipped with an Allegra gradient system and a circularly polarized frequency headcoil. In all imaging experiments, the subject's head was fixated with foam pads to minimize movement during the experiment. A T1-weighted high-resolution data set covering the whole brain was collected for each subject with a three-dimensional magnetization-prepared rapid acquisition gradient echo with $1.2 \times 1 \times 1\ mm^3$ (experiment 1), respectively, a three-dimensional modified driven equilibrium Fourier transform sequence with $1 \times 1 \times 1\ mm^3$ (experiment 2).

To reduce perceptual and physiological interactions of the blood oxygen level-dependent (BOLD) signal caused by the acoustic noise produced by switching magnetic field gradients in fMRI with activity

induced by experimental acoustic stimulation, we used a recently developed novel low-impact noise acquisition fMRI sequence, which increases in the dynamic range of BOLD signal (Seifritz et al., 2006) for functional imaging. In short, this sequence elicits a scanner gradient acoustic noise, which is perceived to be continuous. By way of contrast, conventional echoplanar imaging sequences produce a pulsed scanner noise pattern, which could heavily interfere with the temporal structure of the acoustic stimulation used in our study [for detailed illustration of sound envelopes of scanner noise, see Seifritz et al. (2006), their Fig. 1].

The functional volumes were positioned parallel to the lateral sulcus.

*Imaging parameters of the cross-sectional experiment (1.5 T).* Imaging parameters of the cross-sectional experiment (1.5 T) were as follows: gradient-recalled echoplanar low-impact noise acquisition imaging sequence with 16 image slices having a thickness of 5 mm and a volume repetition time (TR) of 1850 ms (field of view, $180^2$ mm$^2$; matrix, $64^2$ pixels; echo time, 61 ms; flip angle, 90°; bandwidth, 1280 Hz/pixel; slice acquisition time, 116 ms).

*Imaging parameters of the longitudinal experiment (3 T).* Imaging parameters of the longitudinal experiment (3 T) were as follows: gradient-recalled echoplanar low-impact noise acquisition imaging sequence with 20 image slices having a thickness of 4 mm and a volume TR of 1880 ms (field of view, $220^2$ mm$^2$; matrix, $64^2$ pixels; echo time, 30 ms; flip angle, 90°; bandwidth, 1280 Hz/pixel; slice acquisition time, 62 ms).

### *Data preprocessing*

Image time courses were processed using the software package BrainVoyager QX (Brain Innovation): for each subject, the first two echoplanar images were discarded to allow for magnetization signal full saturation, and all the remaining scans were realigned to the first included volume scan using a Levenberg–Marquardt algorithm optimizing three translation and three rotation parameters on a resampled version of each image. The resulting head motion-corrected time series were corrected for the different slice scan times using a cubic spline interpolation procedure and then filtered in the temporal domain. For temporal filtering, a high-pass filter with cutoff to six cycles per time course (114 s) was used to reduce linear and nonlinear trends in the time courses. Using the results of the image registration with three-dimensional anatomical scans, the functional image–time series were warped into Talairach space and resampled into 3 mm isotropic voxel time series. Finally, to perform a group-level analysis, the resampled volume time series were spatially filtered (smoothing) using a 6 mm full-width at half-maximum Gaussian kernel.

### *Statistical analysis*

The variance of all image time series was estimated voxelwise according to a random effects convolution-based general linear model (GLM) analysis (Friston et al., 1995, 1999). Three "event-type" predictors of interest encoding the responses to the three deviant types and one "block-type" predictor of no interest encoding the response to the standard stimulus against a baseline of no auditory stimulation were defined using the double-gamma function (Friston et al., 1998) as hemodynamic input function for the linear convolution. In total, the design matrix included five predictors, three predictors of interest (for the deviant events) and two confounds (for the standard response and the "constant" baseline).

For each subject and each voxel included in the slab of imaging, the five "$\beta$" weights of the five regressors were estimated according to a GLM fit–refit procedure, which ensured a correction of residual serial correlation in the error terms according to a first-order autoregressive model (Bullmore et al., 1996).

To draw population-level inferences from statistical maps, the three $\beta$ estimates for the predictors of interest at each voxel entered a second-level ANOVA with subjects treated as random observations (random-effects ANOVA). Two different factorial designs were defined for the random-effects ANOVA of the cross-sectional and the longitudinal study data. For the cross-sectional experiment, a two-way ANOVA table was prepared, with one within-subject factor for the "temporal novelty" effect (including three levels for the three deviant types) and one between-subject factor for the "musicianship" effect (including two levels for the two groups of professional musicians and nonmusicians). For the longitudinal study, a three-way ANOVA table was prepared, with two within-subject factors for the temporal novelty effect (including three levels for the three deviant types) and for the "musical training" effect (including two levels for pretraining and posttraining measurements) and one between-subject factor for the musicianship effect (including two levels for the two groups of musical and nonmusical students). The resulting $F$ maps for the main effects of the task in all groups and both studies, for the two-way interaction "temporal novelty by musicianship" in the cross-sectional study and "temporal novelty by training" in the longitudinal study and for the three-way interaction "temporal novelty by musical training by musicianship" in the longitudinal study were overlaid on a Montreal Neurological Institute template brain. To localize the significant effects on the average anatomy, a threshold was applied to the $F$ maps, which protected against false-positive clusters at 5% (corrected for multiple comparisons). Starting from the uncorrected threshold of $p = 0.001$, a whole-slab cluster-level correction approach based on Monte Carlo simulations (Forman et al., 1995; Etkin et al., 2004) was used to define the corresponding minimum cluster size to apply. Only clusters that survived this thresholding procedure are reported in Results.

To test for correlations between BOLD responses to temporal novelty and behaviorally assessed musical abilities, a region of interest (ROI) was functionally defined from those voxels that exhibited a statistically significant three-way interaction (temporal novelty by training by musicianship) after the voxel-based analysis of the data from the longitudinal study (experiment 2). To preclude any circularity of data analysis (Kriegeskorte et al., 2009), we then applied this ROI mask derived from experiment 2 (longitudinal study) to the independent fMRI data acquired in experiment 1 (cross-sectional study). The average time courses from these voxels were extracted for each subject (experiment 1) and resubmitted to the same GLM (ROI-GLM). The resulting ROI-GLM fits were finally correlated with individual musical abilities of the subjects as evaluated with the AMMA test (see supplemental Table 1, available at www.jneurosci.org as supplemental material) for the cross-sectional experiment using an analysis of covariance (ANCOVA) model with one categorical factor (musicianship) and one continuous factor (AMMA score).

## Results

### Cross-sectional study (experiment 1)

We presented a temporal mismatch paradigm consisting of regularly spaced sine tones interspersed with infrequently occurring deviant SOAs (Fig. 1) to groups of professional musicians ($n = 7$) and nonmusicians ($n = 7$), matched for age and gender (for detailed subject characterization, see supplemental Table 1, available at www.jneurosci.org as supplemental material), to test for differential brain responses to temporal novelty in the acoustic domain between groups.

The occurrence of standard sine tones, presented with shorter SOAs (142, 130, 100 ms) compared with the standard SOA (150 ms) (Fig. 1), elicited a mismatch response in the temporal plane of the right hemisphere ($p < 0.05$, corrected), which depended on the degree of deviance. The greater the deviance (i.e., the shorter the deviant SOAs), the more neural activity was induced (Fig. 2*A*,*B*). BOLD activity in the right planum temporale induced by a temporal mismatch did not differ significantly between musicians and nonmusicians.

However, a second-level two-way interaction analysis with the factors temporal novelty (three levels of temporal deviance) and musicianship (professional musicians vs nonmusician) revealed enhanced BOLD responses to novelty of temporal structure in musicians' left anterior hippocampus ($p < 0.05$, corrected) (Fig. 2*C*) (for details, see supplemental Fig. S2, available at www.jneurosci.org as supplemental material).

### Longitudinal study (experiment 2)

We presented the temporal mismatch paradigm from experiment 1 to an independent sample of music students and stu-

dents from other faculties ($n = 40$) (supplemental Table 2, available at www.jneurosci.org as supplemental material). To evaluate the impact of academic musical training on temporal novelty detection, BOLD responses to temporal novelty were measured twice in both groups. Music students were examined before they started musical training at university level, and after completing the first two semesters at university, which include aural skill training (see supplemental material, available at www.jneurosci.org). In addition, aural skills were quantified behaviorally using a music dictation, to test for changes across time. We used percentage measures to compare the results of the two music dictations [baseline (1) and follow-up (2)] (see Materials and Methods) of music students. Results showed a high interindividual variance. The students achieved a mean of 65.6% (20.5–99.6; SD, 27.2) correct answers in music dictation 1 (Mozart: symphony KV 319, 2d movement, measures 1–18) and 77.3% (20.7–100; SD, 27.4) in music dictation 2 [J. S. Bach: cantata BWV 119, no. 5 (alto aria), measures 1–13]. Using a Wilcoxon test for nonparametric data revealed a significant time effect (i.e., the music students performed better after one-half year of ear training) ($Z = -2.275$; $p = 0.023$, two-tailed).

Students from other faculties, however, did not receive any aural skill training between the two longitudinal evaluation time points in addition to the formal education specific to their individual fields of study.

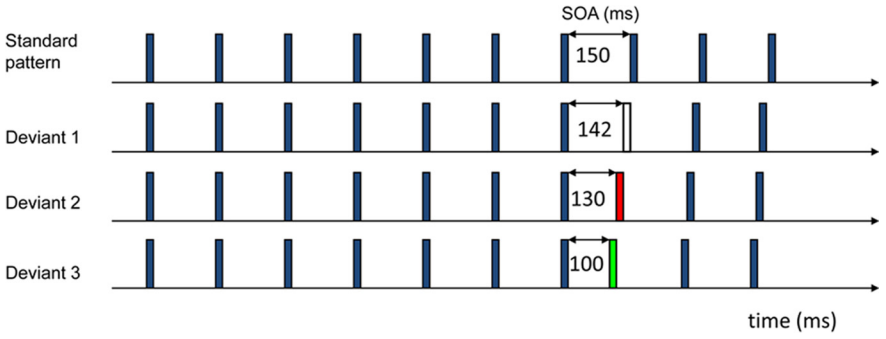


**Figure 1.** Schematic design of stimulation (experiments 1 and 2). Regularly timed sine tones (50 ms duration; 5 ms linear rise and fall times; carrier frequency, 1 kHz) with a standard SOA of 150 ms were presented to the subjects through MR-compatible headphones binaurally. Intermittently, subjects were exposed to interspersed stimuli with a shorter SOA of 142, 130, or 100 ms, respectively. The deviant intervals occurred every 16–20 s in a pseudorandomized order.

### Deviance detection in the planum temporale of the right hemisphere

Computing the main effect of deviant SOAs (142, 130, 100 ms) in the cohort of all students participating in the longitudinal study revealed BOLD activity related to novelty in temporal structure in the planum temporale of the right hemisphere, consistent with the results of the cross-sectional study (Fig. 3*A*). Like in the cross-sectional study, the amplitude of the BOLD response in this region depended on the degree of deviance (i.e., the greater the irregularity compared with the standard temporal pattern, the more neural activity is elicited in this region) (Fig. 3*B*).

### Group differences in temporal novelty detection not related to academic musical training

Evaluating the impact of musicianship (musical students vs students of other faculties) on temporal novelty detection in the longitudinal sample without considering the factor training (i.e., the impact of academic musical training between the two longitudinal scan sessions separated by academic aural skills training) in a two-way interaction analysis (temporal novelty by group) revealed a significant interaction ($p < 0.05$, corrected) in the insula and precuneus of the right hemisphere in response to temporal novelty (supplemental Fig. S1, available at www.jneurosci.org as supplemental material). In these right hemispheric regions, musical students showed enhanced BOLD responses to temporal deviance compared with other students. However, in contrast to the cross-sectional study, this two-way interaction revealed no significant differences in temporal novelty processing in left anterior hippocampus between the groups of the longitudinal sample.

### Effect of academic musical training on hippocampal novelty detection

To test for the effect of academic musical training on brain responses to temporal novelty, we performed a three-way interaction analysis with the factors temporal novelty (three levels of temporal deviance) by musicianship (musical students vs students of other faculties) by training (measurements before and after musical students received general ear training). We found a significant interaction ($p < 0.05$, corrected) of the three factors in the left anterior hippocampus: in comparison with students of nonmusical faculties, musical students showed enhanced hippocampal BOLD responses to temporal novelty only after refining their musical skills through training at academic level (Fig. 3*C*) (for details, see also supplemental Fig. S3, available at www.jneurosci.org as supplemental material). Note that the evaluation of the two-way interaction between the factors temporal novelty and training (without considering the factor musicianship) revealed significant effects in the left inferior frontal gyrus only ($p < 0.05$, corrected), but not in hippocampus. That is, neither the reexposition to the temporal novelty paradigm across groups (factor training) nor the factor musicianship (see above for results of the interaction temporal novelty by musicianship) alone, but only their interaction accounts for the enhanced hippocampal activation, suggesting training-related changes of acoustic novelty detection capabilities in left anterior hippocampus in musical students receiving (task-unrelated) training of musical skills at university.

In summary, like in professional musicians with an academic musical background (see results of cross-sectional study), a general training of aural skills at academic level enhanced BOLD activity in left anterior hippocampus in response to temporal novelty in musical students.

### Behavioral correlates of enhanced hippocampal novelty detection in musicians

We tested for correlations between BOLD responses to temporal novelty in regions showing training-related plasticity (i.e., left hippocampus) and musical aptitude as evaluated using a common standardized measure (AMMA test) (Gordon, 1998). To avoid any circularity of statistical analysis (Kriegeskorte et al., 2009), we used our two independent datasets from experiments 1 and 2 for ROI selection and subsequent analysis. More specifically, first the ROI was functionally defined by the voxel-based three-way interaction analysis with the factors temporal novelty, musicianship, and training in the longitudinal study (experiment 2), representing the brain area in which we observed training-

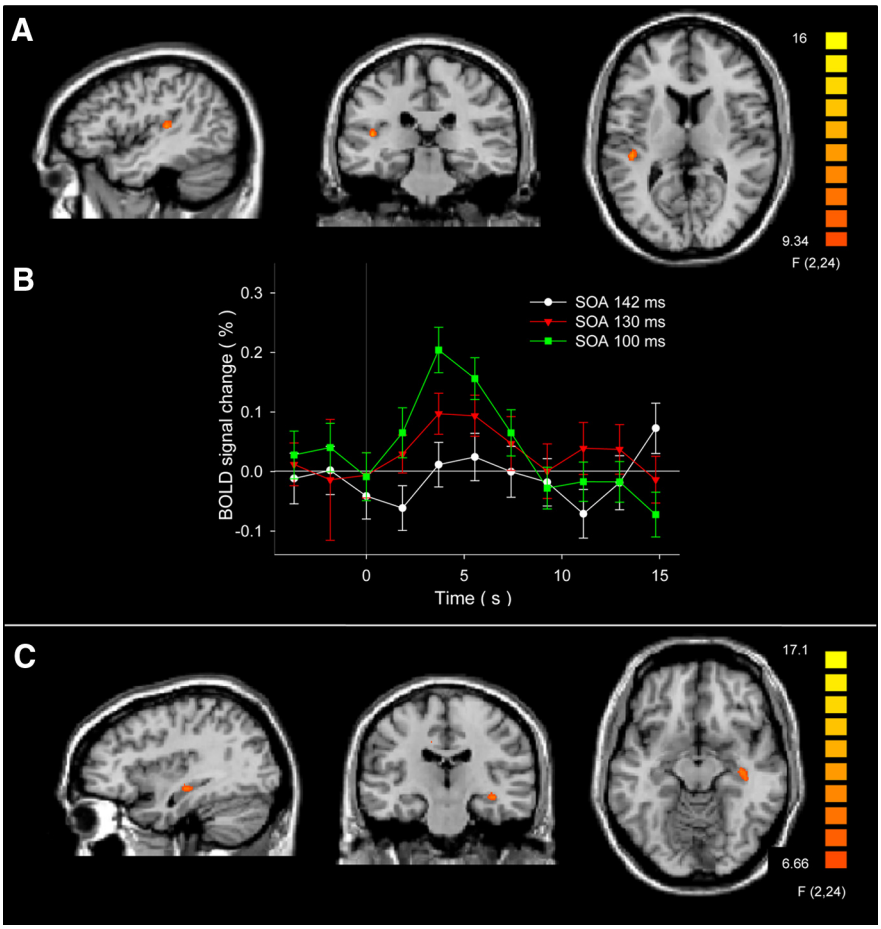


**Figure 2.** BOLD responses to temporal novelty in the cross-sectional population (experiment 1). ***A***, Computing the main effect for all grades of irregular SOAs reveals that deviant events embedded in an otherwise regular temporal pattern induce BOLD activity in planum temporale of the right hemisphere [Talairach coordinates (in mm): $x = +52, y = -27, z = +8$] in musicians and nonmusicians ( $p < 0.05$, cluster level correction). ***B***, Event-related averaging of BOLD signal changes in right temporal plane (rPT) elicited by deviant events (SOA of 142, 130, or 100 ms) shows a parametric variation of the novelty response according to the different degrees of deviance (i.e., the greater the deviance to the regular pattern the more neural activity in rPT is induced). Error bars indicate SEM. ***C***, Second-level two-way interaction analysis with the factors temporal novelty (three levels of temporal deviance) and musicianship (professional musicians vs nonmusicians) revealed enhanced BOLD responses to novelty of temporal structure in musicians' left anterior hippocampus ($x = -35, y = -19, z = -8$; $p < 0.05$, corrected).

related changes (left hippocampus) (Fig. 3). Then, using independent data from subjects of experiment 1, we found that BOLD responses to temporal novelty within this ROI are correlated with the individuals' musical abilities as evaluated separately using the AMMA test (ANCOVA, $F = 4.81$, $p = 0.05$), pointing to the behavioral significance of the functional differences found in musicians' hippocampus.

## Discussion

### Deviance detection in the planum temporale of the right hemisphere

We found that right-hemispheric secondary auditory cortex is crucially involved in detection of temporal pattern variations embedded in an otherwise temporally regular sequence of tones. The planum temporale has been previously suggested to be a "computational hub," which is essential for segregating spectrotemporal sound patterns and matching them with stored representations (Griffiths and Warren, 2002), and mismatches in the temporal domain seem to be preferentially processed in the right hemisphere (Mustovic et al., 2003; Herdener et al., 2007). The modulation of BOLD response amplitude with degree of deviance in this region is consistent with electrophysiological data showing a parametric variation of oddball induced activity with deviance modulation (Näätänen et al., 2001). However, BOLD activity to temporal mismatch in this region does not differ between groups of musicians and nonmusicians, neither in our cross-sectional comparison of professional musicians versus nonmusicians (experiment 1) nor in our longitudinal study evaluating the impact of musical training on acoustic novelty detection (experiment 2). This indicates a comparable detection of temporal novelty in secondary auditory regions of both musicians and nonmusicians.

### Enhanced novelty detection in left anterior hippocampus in musical experts

Significant group differences in the cross-sectional comparison (experiment 1) of temporal novelty detection in the auditory domain were observed in the anterior left hippocampus, which shows enhanced BOLD response to temporal irregularities in professional musicians compared with nonmusicians. Therefore, our data suggest that, although secondary auditory cortex is crucially involved in the preattentive processing of temporal acoustic patterns, differential activation of the left anterior hippocampus represents a functional correlate of musical expertise related to sound pattern processing. This notion is corroborated by the correlation of musical aptitude measures with hippocampal BOLD responses.

The hippocampus has been previously suggested to be involved in the encoding of time intervals (Knight et al., 2004), and in particular anterior regions of the left hippocampus formation in humans have been proposed to index novelty (Strange et al., 1999). Furthermore, the ability to act as a relational operator enabling comparisons between data input and stored data and thereby detecting changes and novel events in the sensory environment has been previously attributed to the hippocampus based on theoretical (Gray and Rawlins, 1986) and animal (Vinogradova, 2001) models and is pivotal when dealing with the recognition of deviant temporal variations embedded in an otherwise regular stimulus pattern. It is also possible that, in analogy to the processing of novelty in the visual domain, the hippocampus is capable of extracting the probabilistic temporal structure of acoustic streams and thus represents the expected information or novelty of any event within a stimulation stream before it occurs (cf. Strange et al., 2005). A recent fMRI study in humans found left hippocampal activation in response to associative novelty, which was investigated by varying the temporal order within a sequence of visual stimuli (Kumaran and Maguire, 2006). Here, we observed that time interval variation within an otherwise regular temporal pattern can induce hippocampal activity. Our data thus support the idea that the hippocampus acts as a novelty detector by compar-

ing actual sensory inputs (e.g., irregularly spaced sine tones) with immediately previous inputs (e.g., preceding periods of acoustic inputs with a regular temporal pattern building up a regular representation or expectation of the acoustic environment), generating novelty signals when previous predictions are violated by sensory reality (Kumaran and Maguire, 2007a), and extend it further to the temporal domain by showing that the hippocampus is also sensitive to variations of time intervals.

### Enhanced hippocampal novelty detection in musicians: talent or training?

The question remains, however, whether exposure (Monaghan et al., 1998) to music, a genetic predisposition (Thompson et al., 2001), or both contribute to the functional differences observed in our cross-sectional study. We thus tested for the impact of musical training at an academic level on hippocampal properties related to temporal novelty detection in a longitudinal approach, and investigated students of the local academy of music before and after receiving ear training during the first semesters of their studies compared with students of other faculties (experiment 2).

First, we specified functional differences in brain responses to temporal novelty between groups of the longitudinal sample without considering training-related effects. We found a significant two-way interaction of the factors temporal novelty by musicianship in the right insular region and the precuneus, pointing to functional differences in these regions between musical and nonmusical students. The right anterior insula has been implicated in the detection of visual and auditory temporal pattern changes (Herdener et al., 2009) and in the processing of novelty tested with various visual, tactile, and auditory oddball paradigms (Linden et al., 1999; Ardekani et al., 2002; Downar et al., 2002). The implication of the right precuneus in novelty detection has also been reported recently (Gur et al., 2007). Functional differences that we found between groups in these regions might be associated with musical talent and/or preacademic musical training. It is important to note, however, that this two-way interaction analysis with the factors temporal novelty by musicianship (and without considering the effect of academic musical training between the two longitudinal sessions) did not reveal any significant functional differences between groups of students ($n = 40$) in the hippocampus. This is in contrast to the results of the cross-sectional study, in which professional musicians showed enhanced hippocampal responses to temporal novelty when compared with nonmusicians ($n = 14$). This suggests that neither the extensive preacademic musical experience of musical students before their education on an academic level nor talent/genetic predisposition is sufficient to account for differential hippocampal novelty detection.

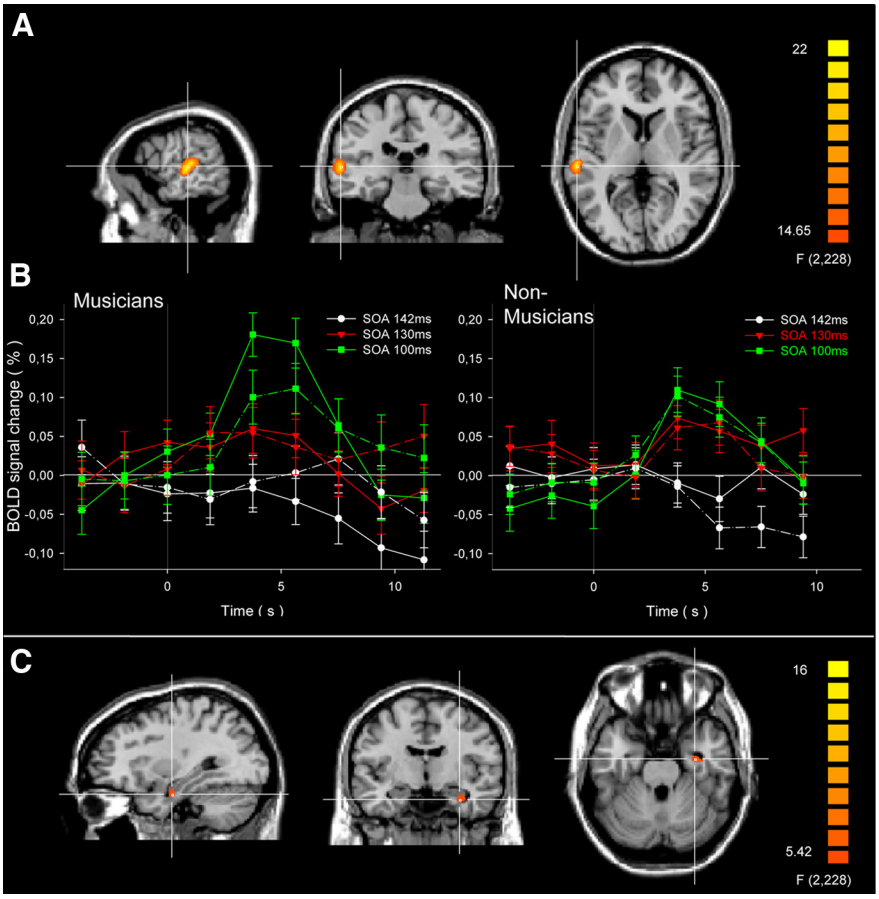


**Figure 3.** BOLD responses to temporal novelty in the longitudinal population (experiment 2). ***A***, Computing the main effect for all grades of irregular SOA reveals BOLD activity increases related to temporal novelty in planum temporale of the right hemisphere ($x = 57, y = -25, z = 7; p < 0.05$, corrected) in students of all faculties. ***B***, Event-related averaging of BOLD signal changes in right temporal plane induced by different temporal irregularities (SOA of 142, 130, or 100 ms) across subjects and across sessions shows parametric variation of novelty responses (solid lines represent values at baseline, and dashed lines at follow up) (compare also with Fig. 2 *B*). Error bars indicate SEM. ***C***, A second-level, three-factorial [temporal novelty (three levels of temporal deviance) by musicianship (musical students vs students of other faculties) by training (measurements before and after musical students received general ear training)], whole-brain-based interaction analysis showing enhanced BOLD responses in left hippocampus ($x = -32, y = -9, z = -22$) to temporal novelty only in musical students after refining of their musical skills by training at an academic level ( $p < 0.05$, corrected).

The activation of the left inferior frontal gyrus in response to temporal novelty as revealed by the two-way interaction analysis with the factors temporal novelty by training is consistent with previous electrophysiological studies showing an involvement of inferior frontal regions in the detection of changes of various acoustic stimulus features (Deouell et al., 1998; Rinne et al., 2000; Opitz et al., 2002). More specifically, this region has been suggested to be especially sensitive to variations of sounds in the temporal domain (Molholm et al., 2005), and our data thus support the notion of inferior frontal gyrus as a detector of temporal acoustic oddballs. However, activity in this region did not differ between groups.

In a next step, we specified the effects of academic musical training on brain responses to temporal novelty in a whole-brain interaction analysis of the longitudinal data with the three factors temporal novelty by musicianship by training. Considering training-related plasticity in students, we found a significant, whole-brain-corrected three-way interaction ( $p < 0.05$, corrected) in left anterior hippocampus (but not in any other brain

region) (Fig. 3*C*) with enhanced BOLD activity in musical students after they had received academic ear training. This finding demonstrates that musical students show enhanced BOLD activity to temporal novelty in the same region (i.e., left anterior hippocampus) where we found functional differences between professional musicians, who were all experts in ear training, and nonmusicians only after receiving academic ear training, suggesting that we were able to identify functional correlates of training-induced changes in sound pattern processing.

In addition, we found a correlation of hippocampal sensitivity to temporal novelty with musical abilities. Based on our imaging and behavioral data, we assume that the observed changes in hippocampal activity in musicians represent a functional correlate of a tuning of aural skills related to time interval perception during the course of their studies. There is also evidence showing that lesions of the left hippocampus in humans impair performance related to time interval discrimination within otherwise regular acoustic streams similar to the sounds presented in our study (Samson et al., 2001), supporting the behavioral relevance of the left hippocampus for temporal novelty detection in the acoustic modality.

Taken together, our data extend the notion of the hippocampus as a novelty detector to the temporal domain of the acoustic modality and expand previous findings on the impact of musical expertise on brain functions (Elbert et al., 1995; Pantev et al., 1998; Schneider et al., 2002; Parbery-Clark et al., 2009) to the hippocampal region. Moreover, to our knowledge, this is the first longitudinal investigation showing that training can induce functional plasticity in the adult human hippocampus, as exemplified using musicians' brains as a model.

## References


Ardekani BA, Choi SJ, Hossein-Zadeh GA, Porjesz B, Tanabe JL, Lim KO, Bilder R, Helpern JA, Begleiter H (2002) Functional magnetic resonance imaging of brain activity in the visual oddball task. Brain Res Cogn Brain Res 14:347–356.

Brown MW, Bashir ZI (2002) Evidence concerning how neurons of the perirhinal cortex may effect familiarity discrimination. Philos Trans R Soc Lond B Biol Sci 357:1083–1095.

Bullmore E, Brammer M, Williams SC, Rabe-Hesketh S, Janot N, David A, Mellers J, Howard R, Sham P (1996) Statistical methods of estimation and inference for functional MR image analysis. Magn Reson Med 35:261–277.

Deouell LY, Bentin S, Giard MH (1998) Mismatch negativity in dichotic listening: evidence for interhemispheric differences and multiple generators. Psychophysiology 35:355–365.

Downar J, Crawley AP, Mikulis DJ, Davis KD (2002) A cortical network sensitive to stimulus salience in a neutral behavioral context across multiple sensory modalities. J Neurophysiol 87:615–620.

Draganski B, Gaser C, Kempermann G, Kuhn HG, Winkler J, Büchel C, May A (2006) Temporal and spatial dynamics of brain structure changes during extensive learning. J Neurosci 26:6314–6317.

Ekstrom AD, Kahana MJ, Caplan JB, Fields TA, Isham EA, Newman EL, Fried I (2003) Cellular networks underlying human spatial navigation. Nature 425:184–188.

Elbert T, Pantev C, Wienbruch C, Rockstroh B, Taub E (1995) Increased cortical representation of the fingers of the left hand in string players. Science 270:305–307.

Etkin A, Klemenhagen KC, Dudman JT, Rogan MT, Hen R, Kandel ER, Hirsch J (2004) Individual differences in trait anxiety predict the response of the basolateral amygdala to unconsciously processed fearful faces. Neuron 44:1043–1055.

Forman SD, Cohen JD, Fitzgerald M, Eddy WF, Mintun MA, Noll DC (1995) Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. Magn Reson Med 33:636–647.

Friston KJ, Holmes AP, Poline JB, Grasby PJ, Williams SC, Frackowiak RS, Turner R (1995) Analysis of fMRI time-series revisited. Neuroimage 2:45–53.

Friston KJ, Josephs O, Rees G, Turner R (1998) Nonlinear event-related responses in fMRI. Magn Reson Med 39:41–52.

Friston KJ, Holmes AP, Worsley KJ (1999) How many subjects constitute a study? Neuroimage 10:1–5.

Gordon EE (1998) Introduction to research and psychology of music. Chicago: GIA.

Gray JA, Rawlins JNP (1986) Comparator and buffer memory: an attempt to integrate two models of hippocampal function. In: The hippocampus, pp 159–201. New York: Plenum.

Griffiths TD, Warren JD (2002) The planum temporale as a computational hub. Trends Neurosci 25:348–353.

Gur RC, Turetsky BI, Loughead J, Waxman J, Snyder W, Ragland JD, Elliott MA, Bilker WB, Arnold SE, Gur RE (2007) Hemodynamic responses in neural circuitries for detection of visual target and novelty: an event-related fMRI study. Hum Brain Mapp 28:263–274.

Herdener M, Esposito F, Di Salle F, Lehmann C, Bach DR, Scheffler K, Seifritz E (2007) BOLD correlates of edge detection in human auditory cortex. Neuroimage 36:194–201.

Herdener M, Lehmann C, Esposito F, di Salle F, Federspiel A, Bach DR, Scheffler K, Seifritz E (2009) Brain responses to auditory and visual stimulus offset: shared representations of temporal edges. Hum Brain Mapp 30:725–733.

Kempermann G, Kuhn HG, Gage FH (1997) More hippocampal neurons in adult mice living in an enriched environment. Nature 386:493–495.

Knight DC, Cheng DT, Smith CN, Stein EA, Helmstetter FJ (2004) Neural substrates mediating human delay and trace fear conditioning. J Neurosci 24:218–228.

Knight R (1996) Contribution of human hippocampal region to novelty detection. Nature 383:256–259.

Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI (2009) Circular analysis in systems neuroscience: the dangers of double dipping. Nat Neurosci 12:535–540.

Kumaran D, Maguire EA (2006) An unexpected sequence of events: mismatch detection in the human hippocampus. PLoS Biol 4:e424.

Kumaran D, Maguire EA (2007a) Which computational mechanisms operate in the hippocampus during novelty detection? Hippocampus 17:735–748.

Kumaran D, Maguire EA (2007b) Match mismatch processes underlie human hippocampal responses to associative novelty. J Neurosci 27:8517–8524.

Linden DE, Prvulovic D, Formisano E, Völlinger M, Zanella FE, Goebel R, Dierks T (1999) The functional neuroanatomy of target detection: an fMRI study of visual and auditory oddball tasks. Cereb Cortex 9:815–823.

Lledo PM, Alonso M, Grubb MS (2006) Adult neurogenesis and functional plasticity in neuronal circuits. Nat Rev Neurosci 7:179–193.

Maguire EA (2001) Neuroimaging, memory and the human hippocampus. Rev Neurol (Paris) 157:791–794.

Maguire EA, Gadian DG, Johnsrude IS, Good CD, Ashburner J, Frackowiak RS, Frith CD (2000) Navigation-related structural change in the hippocampi of taxi drivers. Proc Natl Acad Sci U S A 97:4398–4403.

Molholm S, Martinez A, Ritter W, Javitt DC, Foxe JJ (2005) The neural circuitry of pre-attentive auditory change-detection: an fMRI study of pitch and duration mismatch negativity generators. Cereb Cortex 15:545–551.

Monaghan P, Metcalfe NB, Ruxton GD (1998) Does practice shape the brain? Nature 394:434.

Munte TF, Altenmuller E, Jancke L (2002) The musician's brain as a model of neuroplasticity. Nat Neurosci 3:473–478.

Mustovic H, Scheffler K, Di Salle F, Esposito F, Neuhoff JG, Hennig J, Seifritz E (2003) Temporal integration of sequential auditory events: silent period in sound pattern activates human planum temporale. Neuroimage 20:429–434.

Näätänen R, Escera C (2000) Mismatch negativity: clinical and other applications. Audiol Neurootol 5:105–110.

Näätänen R, Tervaniemi M, Sussman E, Paavilainen P, Winkler I (2001) "Primitive intelligence" in the auditory cortex. Trends Neurosci 24:283–288.

Opitz B, Rinne T, Mecklinger A, von Cramon DY, Schröger E (2002) Dif-

ferential contribution of frontal and temporal cortices to auditory change detection: fMRI and ERP results. Neuroimage 15:167–174.

Pantev C, Oostenveld R, Engelien A, Ross B, Roberts LE, Hoke M (1998) Increased auditory cortical representation in musicians. Nature 392: 811–814.

Parbery-Clark A, Skoe E, Kraus N (2009) Musical experience limits the degradative effects of background noise on the neural processing of sound. J Neurosci 29:14100–14107.

Peretz I (2006) The nature of music from a biological perspective. Cognition 100:1–32.

Rinne T, Alho K, Ilmoniemi RJ, Virtanen J, Näätänen R (2000) Separate time behaviors of the temporal and frontal mismatch negativity sources. Neuroimage 12:14–19.

Rüsseler J, Altenmüller E, Nager W, Kohlmetz C, Münte TF (2001) Event-related brain potentials to sound omissions differ in musicians and non-musicians. Neurosci Lett 308:33–36.

Samson S, Ehrle N, Baulac M (2001) Cerebral substrates for musical temporal processes. Ann N Y Acad Sci 930:166–178.

Schneider P, Scherg M, Dosch HG, Specht HJ, Gutschalk A, Rupp A (2002) Morphology of Heschl's gyrus reflects enhanced activation in the auditory cortex of musicians. Nat Neurosci 5:688–694.

Seifritz E, Di Salle F, Esposito F, Herdener M, Neuhoff JG, Scheffler K (2006) Enhancing BOLD response in the auditory system by neurophysiologically tuned fMRI sequence. Neuroimage 29:1013–1022.

Sokolov EN (1963) Higher nervous functions; the orienting reflex. Annu Rev Physiol 25:545–580.

Strange BA, Dolan RJ (2001) Adaptive anterior hippocampal responses to oddball stimuli. Hippocampus 11:690–698.

Strange BA, Fletcher PC, Henson RN, Friston KJ, Dolan RJ (1999) Segregating the functions of human hippocampus. Proc Natl Acad Sci U S A 96:4034–4039.

Strange BA, Duggins A, Penny W, Dolan RJ, Friston KJ (2005) Information theory, novelty and hippocampal responses: unpredicted or unpredictable? Neural Netw 18:225–230.

Thompson PM, Cannon TD, Narr KL, van Erp T, Poutanen VP, Huttunen M, Lönnqvist J, Standertskjöld-Nordenstam CG, Kaprio J, Khaledy M, Dail R, Zoumalan CI, Toga AW (2001) Genetic influences on brain structure. Nat Neurosci 4:1253–1258.

Vinogradova OS (2001) Hippocampus as comparator: role of the two input and two output systems of the hippocampus in selection and registration of information. Hippocampus 11:578–598.

Yamaguchi S, Hale LA, D'Esposito M, Knight RT (2004) Rapid prefrontal-hippocampal habituation to novel events. J Neurosci 24: 5356–5363.

Mem Cogn (2012) 40:514–527
DOI 10.3758/s13421-011-0167-z

# Making related errors facilitates learning, but learners do not know it

**Barbie J. Huelser · Janet Metcalfe**

Published online: 9 December 2011


**Abstract** Producing an error, so long as it is followed by corrective feedback, has been shown to result in better retention of the correct answers than does simply studying the correct answers from the outset. The reasons for this surprising finding, however, have not been investigated. Our hypothesis was that the effect might occur only when the errors produced were related to the targeted correct response. In Experiment 1, participants studied either related or unrelated word pairs, manipulated between participants. Participants either were given the cue and target to study for 5 or 10 s or generated an error in response to the cue for the first 5 s before receiving the correct answer for the final 5 s. When the cues and targets were related, error-generation led to the highest correct retention. However, consistent with the hypothesis, no benefit was derived from generating an error when the cue and target were unrelated. Latent semantic analysis revealed that the errors generated in the related condition were related to the target, whereas they were not related to the target in the unrelated condition. Experiment 2 replicated these findings in a within-participants design. We found, additionally, that people did not know that generating an error enhanced memory, even after they had just completed the task that produced substantial benefits.

**Keywords** Memory · Errors · Generation · Metacognition · Associative learning

B. J. Huelser (✉) · J. Metcalfe
Department of Psychology, Columbia University,
401 Schermerhorn Hall, 1190 Amsterdam Ave., MC 5501,
New York, NY 10027, USA
e-mail: bjh2135@columbia.edu

This article addresses the effect of making errors on learning. Should one learn by studying materials without making mistakes or by attempting to produce answers and committing inevitable errors that such attempts entail? When errors are left uncorrected, they typically remain incorrect (Butler, Karpicke & Roediger, 2008; Fazio, Huelser, Johnson & Marsh, 2010; Metcalfe & Kornell, 2007; Pashler, Cepeda, Wixted & Rohrer, 2005; Pashler, Zarow & Triplett, 2003). However, feedback is highly effective in allowing the learner to correct previously incorrect answers (Butler et al., 2008; Metcalfe, Kornell & Finn 2009; Pashler et al., 2005; Pashler et al., 2003). In this article, only errors followed by corrective feedback were considered. The question here was whether, and under what conditions, committing an error facilitates learning. Although the main focus of this article is the memorial consequences for errorful, as compared with errorless, learning, a related question of interest is the following: Are learners *aware* of the circumstances in which committing errors can be effective for improving learning? Accurate metacognitive knowledge is important for metacognitive control and strategy selection (Kornell & Son, 2009; Metcalfe & Finn, 2008). If the learner is not aware of the potential efficacy of a learning strategy, he or she might implement suboptimal strategies. Hence, people's metacognitions about the effects of errors may be nearly as important as the effects of the errors themselves.

From a theoretical standpoint, there is reason to believe that even corrected errors might impede learning. An error, in essence, is often thought to be conflicting or competing information with regard to the correct response. As such, it should create an interference situation. In standard proactive interference paradigms, the first pairing of a target (B)

with a particular cue (A) results in interference when cue A is later paired with a different response (C) (J. R. Anderson & Reder, 1999; M. C. Anderson & Neely, 1996; Barnes & Underwood, 1959; Loftus, 1979; McGeoch, 1942; Melton & Irwin, 1940; Osgood, 1949; Webb, 1917). Although there are several theories concerning how this interference arises (e.g., J. A. Anderson, 1973; J. R. Anderson & Bower, 1972; Eich, 1982; Gillund & Shiffrin, 1984; Hintzman, 1984; Metcalfe, 1990; Osgood, 1949), there is general agreement that it does occur. Interference from errors might be expected to be even greater than interference theory would normally predict, since interference theory does not take into account whether or not the interfering information is self-produced. Incorrect information that is self-generated might be even more difficult to overcome than a provided response, because the process of self-generation has been shown to enhance memory for the response (Slamecka & Graf, 1978; for reviews, see Bertsch, Pesta, Wiscott & McDaniel, 2007; Mulligan & Lozito, 2005).

In accordance with the rationale described above, it has sometimes been recommended that errors be eliminated during learning (Glaser, 1990). For example, Guthrie (1952) suggested that errors should be avoided because, when errors are practiced, the incorrect response to a particular stimulus will be strengthened. Furthermore, errorless, as compared with trial-and-error, learning has been shown to be beneficial for people with memory impairments, including Alzheimer's disease, schizophrenia, Korsokoff's syndrome, and trauma (see Clare & Jones, 2008, for a review). One concern with generalizing from this line of empirical research, however, is that the benefits of errorless over errorful learning have been found primarily in patient populations and may not apply to typical learners. Nevertheless, in an experiment by Cunningham and Anderson (1968), worse retention was found after typical participants had been forced to guess, rather than following a simple presentation of the to-be-remembered material.

Despite the arguments that the generation of errors impedes learning, several researchers have found that error-generation is not detrimental to memory of subsequently learned correct answers. One way of examining the effect of errors on learning is by forcing responses for every item on a test, as compared with leaving participants free to answer only when they so choose. Forced responding results in more errors than does free responding. However, on a later test of definition terms, using this procedure with both college undergraduates and 6th grade students, Metcalfe and Kornell (2007) found neither benefit nor impairment for forced, as compared with free, responding. Similarly, Kang et al. (2011) found that forced guessing did not lead to better or worse memory for the correct answer on a later retention test, neither immediately nor at a 1-week delay. However, it is impossible to know whether the lack of a difference might have occurred because people in the free-responding condition generated errors to the same extent as people in the forced-responding condition, but did not overtly express them. It is also not known what kinds of errors were produced under the forced-guessing procedures and, in particular, whether they were related or unrelated to the targets. Research based on multiple-choice quizzing prior to learning a lesson in a classroom setting also suggests that pretesting—which results in many errors—neither helps nor hurts memory for the correct information (McDaniel, Agarwal, Huelser, McDermott & Roediger, 2011). No difference in memory was found for items quizzed on a pretest, as compared with nonquizzed items.

In contrast to the findings above, however, there are some studies showing that under certain circumstances, making errors helps learning. Richland, Kornell and Kao (2009) found enhanced memory for material from reading passages when the to-be-remembered material was tested using cued recall questions prior to the reading of the passages, even though participants did not answer these pretest questions correctly. Izawa (1967, 1970) has also shown that multiple incorrect retrieval attempts enhanced learning: Producing more incorrect responses before receiving feedback led to better memory for the correct feedback than did producing fewer incorrect responses. Parlow and Berlyne (1971) found that participants were better at learning the correct translations for foreign language words when they had previously made an erroneous guess, as compared with when they were exposed to the guesses of others. Kane and Anderson (1978) showed that generating the last word of the sentence, even if it was incorrect, led to enhanced performance over simply reading the sentence. Slamecka and Fevreiski (1983) reported a benefit, above just reading the answer, from trying unsuccessfully to generate it.

Finally, in a paradigm that we will investigate here, Kornell, Hays and Bjork (2009) demonstrated a considerable benefit of prior incorrect guessing for subsequent learning of the correct answer. Participants learned weakly associated word pairs (e.g., *whale–mammal*, *swing–tree*, *together–love*) for a later cued recall test. During the initial learning phase, participants randomly studied word pairs in either a reading mode or an error-generating mode. In the reading mode, both the cue and the target were displayed on the screen for a fixed amount of time (either 5 or 13 s). In the error-generating mode, participants saw only the first word (the cue) for 8 s and had to type a guess into the computer as to what they thought the target would be, followed by the correct cue–target pairing displayed for 5 s. At test, given the cue, participants were required to produce

the correct target and not the original error. Error-generation led to enhanced retention, as compared with both reading conditions.

In sum, it is unclear whether errors during learning hinder, enhance, or simply have no effect on learning. Any of these three options might be possible under different conditions, but it is not yet known what those conditions might be. However, studies in which there was a benefit of error-generation used cue–target pairs that generally seemed to be meaningfully related. For example, the experiments in Kornell et al.'s (2009) study showing beneficial effects used to-be-remembered materials that were weakly associated word pairs. By extension, it might be plausible that errors generated in response to these cues might also have been related, rather than unrelated, to the targets. However, in one of Kornell et al.'s (2009) experiments, no benefit for error-generation was found. In this case, participants guessed answers to fictional general knowledge questions (Berger, Hall & Bahrick, 1999) to which they could not possibly have known anything about the correct answers, such as "What is the last name of the person who invented maladaptibility?" It is likely that the errors that people generated in this particular case were unrelated to the targets. Additional support for the idea that the relatedness of the errors might matter comes from Slamecka and Fevreiski (1983), who compared a generation-followed-by-feedback condition with a read condition. Judges retrospectively evaluated the relatedness of the errors of commission that participants had made, dividing them into those that were related and unrelated to the target. Related errors led to fairly high later recall, whereas unrelated errors and omissions led to low recall. These results suggest that the relatedness of the errors to the target may be an important factor in determining whether errors help or hurt recall—a possibility that was investigated in the experiments that follow.

Finally, given that there is a conflict concerning the effects of errors in the research literature, it is plausible to suppose that the learners themselves might not know whether errors help or hurt learning. As well as exploring the conditions under which errors promote and hinder learning, we also investigated whether, in retrospect, participants were able to accurately monitor whether generating errors helped or hurt their performance on the final test. This question is important, since metacognitive monitoring has been shown to have consequences for strategy selection, referred to as *metacognitive control* (Metcalfe & Finn, 2008; Thiede, Anderson & Therriault, 2003).

## Experiment 1

In Experiment 1, we extended Kornell et al.'s (2009) experiment by contrasting memory for weakly associated word pairs, for which they found the beneficial effect of error-generation, to unrelated word pairs, for which we hypothesized that the effect would not be found. Participants studied word pairs in an error-generation condition and two different read conditions. Half of the participants studied weakly related word pairs, while the other half studied unrelated word pairs. We also tested participants' retrospective metacognitions about their memory performance.

### Method

*Participants* Seventy-one Columbia undergraduates participated for partial fulfillment of a class requirement. Data from 11 nonnative speakers were excluded, leaving data from 60 participants. Mean age was 21.8 years ($SD$ = 6.2), and 68.3% of the participants were female. All participants in both experiments were treated in accordance with APA ethical guidelines.

*Design and materials* The semantic relation of the to-be-remembered materials was manipulated between participants, while learning condition was a within-participants variable, resulting in a 2 (materials: related, unrelated) × 3 (learning condition: read-short, read-long, error-generate) mixed design.[1]

For the related materials condition, 90 weakly associated word pairs were selected from Nelson, McEvoy and Schreiber (1998) norms, closely following Kornell et al.'s (2009) word pair selection criteria. Given the first word, approximately 5% of participants in Nelson et al.'s experiment produced the target as the first associate. Specifically, forward associative strength was between .05 and .054, and backward associative strength was 0. Each word was a minimum of four letters long. For the unrelated materials condition, new materials were selected because in a pilot experiment, cued recall performance was at floor for random word pairs created from the Nelson et al. (1998) norms. Therefore, unrelated word pairs were created from Pavio, Yuille and Madigan (1968) norms. One hundred eighty words were selected (to create 90 word pairs) with relatively high concreteness ratings (6.38–7 on a 1–7 scale) and were a minimum of four letters long. Words were randomly assigned as cues or targets, and three independent coders checked that the so-constructed list of 90 unrelated word pairs contained no accidentally related word pairs.

[1] Because we did not know whether our experiment would replicate the findings of Kornell et al. (2009), we assigned the related materials condition to the first 18 participants, a condition that is most similar to their experiment. After the first set of data on 18 participants was collected in 3 days and it was clear that we were replicating the earlier results, we randomly assigned participants to both materials conditions, beginning the following week.

Mean concreteness ratings were the same for the words assigned as cues and targets ($M$ = 6.77). For each of the between-participants conditions, the 90 word pairs were randomized into three sets of 30 items, which were rotated through each of the study conditions, creating three unique counterbalanced conditions.

*Procedure* This experiment had four phases: learning, distractor, final test, and metacognitive judgment. During the learning phase, 30 word pairs were presented in each of the three conditions (90 word pairs in total). Word pairs were presented in a random order by MediaLab and DirectRT software (Jarvis, 2004). In the error-generation condition, participants were given only the first word (cue) of a word pair, with a text box displayed below. Participants were instructed to think of what the second word might be and to type their response into the text box as quickly as possible. After 5 s, the text box disappeared, and the correct cue–target pairing appeared, with both the cue and the target remaining on the screen for 5 s. In the read-short condition, both the cue and the target were presented together on the screen for 5 s, while in the read-long condition, both the cue and target were presented for 10 s. These conditions were presented in a random order (not blocked). The computer made a soft clicking sound to alert the participants to the presentation of the next word pair. Before the study phase began, participants read instructions on a computer screen. The experimenter also discussed the procedure verbally and ensured that participants understood the task before proceeding. During the instructions, the experimenter expressed that it was extremely difficult to correctly guess the correct target word, to prevent the participants from being discouraged by poor performance on the task. They were instructed to remember the target answer presented by the computer for the later memory test, not the word they had produced. During the distractor phase, participants played a visuospatial computer game for 6 min before continuing to the final test.

The final test was self-paced and consisted of all 90 word pairs presented during the learning phase. For each word pair, the cue was displayed on the screen, with a textbox below. Participants were instructed to type in the correct target for each cue and to provide a guess if unsure of the correct answer. The order of presentation was randomized.

Following the final test, participants made a metacognitive judgment of their performance on the final test on the basis of the initial learning conditions. Instructions were as follows:

> There were three conditions in this experiment: A) together–short: both words displayed on the screen for 5 s, B) together –long: both words displayed on the screen for 10s; C) separate: the first word presented separately (5 s) before both words were displayed (5 s). Which condition helped you learn the word pairs the best for the final test? Please order the conditions in order from which condition led to the BEST to WORST memory on the final test.

Participants subjectively ranked the conditions by entering the associated letter from the best to the worst for memory. We avoided the word *error* in the error-generation condition because we thought that its negative connotation might bias the judgment. Following a demographic questionnaire, all participants were thanked and debriefed.

## Results

Two coders checked for and corrected spelling and typographical mistakes on the original and final tests before analysis of the data. A strict coding rule was followed in which, if the tense (i.e., *clean* vs. *cleaned*/*cleaning*) or form of speech (*dust* vs. *dusty*) was different from the target, that item was coded as incorrect. However, in the few instances in which an item was made plural (*reptile* vs. *reptiles*), it was coded as correct.

*Learning phase performance* Participants in the related materials condition guessed correctly on 3% of the error-generation trials ($SD$ = .03), while no participant in the unrelated materials condition ever correctly guessed the target word during the learning phase ($M$ = .00, $SD$ = .00). All further results reported for the error-generation condition are only from items that were initially answered incorrectly during the learning phase—that is, 97% of the trials for the related materials condition, and 100% of the trials for the unrelated materials condition.[2]

*Final cued recall test correct performance* As is shown in Fig. 1, correct final performance was higher for related materials ($M$ = .64, $SD$ = .19) than for unrelated materials ($M$ = .21, $SD$ = .15), $F(1, 58) = 91.34$, $MSE$ = .09, $p < .001$, $\eta_p^2$ = .61. There was a main effect of learning condition: Error-generation led to the highest proportion correct on the cued recall test, $F(2, 116) = 13.71$, $MSE$ = .01, $p < .001$, $\eta_p^2$ = .19. However, this main effect was qualified by an

[2] Of these errors, 90% were errors of commission for related materials, and 91% for unrelated materials, $t < 1$. Reported data include errors of omission as well, since correct performance on the final cued recall test was not statistically different as a function of prior error type, $F < 1$. For Experiment 2, 96% of errors were errors of commission, and a similar pattern of results for final test performance as a function of prior error type was found. Therefore, results are not conditionalized upon error type, with the exception of latent semantic analysis as it could be computed only for generated errors.

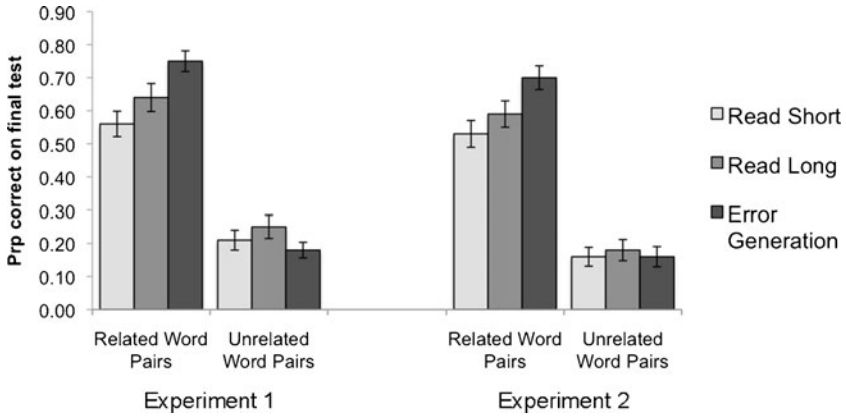


**Fig. 1** Cued Recall Performance. Correct performance on final cued recall test as a function of Learning condition and Materials for both Experiment 1 (between-subjects) and Experiment 2 (within-subjects)

interaction with type of materials. Although error-generation enhanced retention for related materials, it did not enhance performance for unrelated materials, $F(2, 116) = 32.21$, $MSE = .01$, $p < .001$, $\eta_p^2 = .36$. Within related materials, the error-generation condition led to the highest proportion correct on the cued recall test ($M = .74$, $SD = .17$), which was much higher than recall in the read-long condition ($M = .62$, $SD = .23$), $t(29) = 5.14$, $SE = .02$, $p < .001$. The read-short condition led to the lowest proportion correct ($M = .54$, $SD = .21$), which was significantly lower than performance in the read-long condition, $t(29) = 3.54$, $SE = .02$, $p < .01$, and the error-generation condition, $t(29) = 7.37$, $SE = .03$, $p < .001$. With unrelated items, however, the read-long condition led to the highest correct performance ($M = .25$, $SD = .19$), which was significantly better than performance for both the read-short condition ($M = .21$, $SD = .16$), $t(29)$ 2.09, $SE = .02$, $p < .05$, and the error-generation condition ($M = .17$, $SD = .12$), $t(29) = 3.26$, $SE = .02$, $p < .01$. Although the trend favored the read-short condition over error-generation, performance between these two conditions was not significantly different from one another, $t(29) = 1.89$, $SE = .02$, $p = .068$.

*Reaction times* Reaction time (RT) data on the final test were analyzed as a function of accuracy on the final test (correct vs. incorrect), learning condition (read-short, read-long, error-generation), and materials (related, unrelated; see Table 1 for means). RT data are reported in the present section for completeness but will be discussed only in the General Discussion section. Several participants did not have data in all cells in the RT data, and as a result, the degrees of freedom in the analyses given below differ from those given in the basic data for this experiment.

Overall, correct responses ($M = 4.44$ s, $SD = 1.13$) were faster than incorrect responses ($M = 8.63$ s, $SD = 4.06$), $F(1, 54) = 67.65$, $MSE = 21.84$, $p < .001$, $\eta_p^2 = .56$. Collapsed over accuracy, participants were slowest to respond to items on which they had previously generated an error ($M = 7.35$ s, $SD = 3.43$), in comparison with the read-short items ($M = 6.20$ s, $SD = 2.68$) and read-long items ($M = 6.04$ s, $SD = 2.40$), $F(2, 108) = 13.38$, $MSE = 4.20$, $p < .001$, $\eta_p^2 = .20$. There was an interaction between accuracy and learning condition, whereby the difference in RT between items answered incorrectly and correctly on the final test was larger in the error-generation conditions than in the read conditions, $F(2, 108) = 6.30$, $MSE = 3.79$, $p < .01$, $\eta_p^2 = .10$.

Lastly, the relatedness of the materials did not result in differences in RTs. Response latencies were similar regardless of materials condition. There was no difference between related and unrelated materials, $F = 1.06$, $\eta_p^2 = .02$, and materials did not interact with any other factor.

*Error persistence* In the error-generation conditions, more of the initially incorrect responses intruded on the final test for unrelated materials ($M = .20$, $SD = .20$), as compared with related materials ($M = .05$, $SD = .06$), $t(58) = 4.05$, $SE = .04$, $p < .001$.

*Metacognition* Data from 52 participants were included in the metacognitive analyses: 26 from the related materials condition and 26 from the unrelated condition. Exclusions were due to participant failure to assign a distinct metacognitive ranking to each of the three learning conditions. In order to compare performance and metacognitive rankings for each participant, the three conditions were assigned a value on a 0 to 2 scale. The learning condition in which the participant performed best on the final test was assigned a 2; the condition in which he or she performed second best was assigned a 1; the worst was given a score of 0. The same assignment was done for individuals' metacognitive ratings of the three learning conditions.

**Table 1** Mean reaction time in seconds (s) for responding on the final test as a function of Learning condition, Material condition, and Accuracy on the final cued recall test. Standard deviations are provided in parentheses

| | Correct on Final Test | | | Incorrect on Final Test | | |
|---|---|---|---|---|---|---|
| | Read-Short | Read-Long | Error-Generate | Read-Short | Read-Long | Error-Generate |
| *Experiment 1* | | | | | | |
| Related | 3.86 (1.01) | 3.93 (0.78) | 4.27 (0.92) | 7.83 (4.79) | 7.26 (3.43) | 10.10 (5.26) |
| Unrelated | 4.85 (2.15) | 4.57 (1.09) | 5.14 (1.99) | 8.27 (4.06) | 8.42 (4.26) | 9.87 (5.78) |
| *Experiment 2* | | | | | | |
| Related | 3.58 (1.25) | 3.77 (0.92) | 4.19 (1.33) | 6.72 (2.96) | 6.93 (2.96) | 8.00 (3.830) |
| Unrelated | 4.81 (1.83) | 3.82 (1.24) | 4.07 (1.46) | 6.66 (2.71) | 6.44 (2.16) | 8.25 (2.91) |

As can be seen in Fig. 2, participants believed that they performed the best in the read-long condition. They also believed that they had done poorly in both the error-generation condition and read-short condition—regardless of whether the materials were related or unrelated pairs. For the unrelated materials, these metacognitive rankings were approximately correct. However, the participants' beliefs were radically wrong for the related materials: They failed to realize that generating errors greatly facilitated recall under this condition, even after having just experienced the enhanced test performance.

To assess this pattern statistically, metacognitive mean ranking was contrasted with performance mean rankings within each learning condition. These comparisons were done separately for each of the two materials conditions, using the Wilcoxon nonparametric test in lieu of the standard paired-samples *t*-test. Rankings for performance and metacognitive judgments (within materials condition) are not independent, so these contrasts could not be computed. First, for the items in the read-short condition for related materials, there was a trend for actual performance ($M = 0.35$, $SD = 0.56$) to be worse than subjectively reported ($M = 0.65$, $SD = 0.70$), $z = 1.86$, $p = .06$. For the read-long condition, the mean metacognitive ranking was higher ($M = 1.50$, $SD = .65$) than the actual performance ranking ($M = 0.92$, $SD = 0.61$), $z = 2.78$, $p < .01$. Most interesting, however, in the error-generation condition, participants mistakenly believed that their performance was very low ($M = 0.85$ $SD = 0.88$) when it was actually high ($M = 1.73$, $SD = 0.55$), $z = 3.45$, $p < .01$. Within unrelated materials, participants' retrospective metacognitive rankings were very close to actual performance rankings: There was no difference in mean subjective metacognitive ranking, as compared with actual performance rank, for any of the comparisons, $z$s < 1.16.

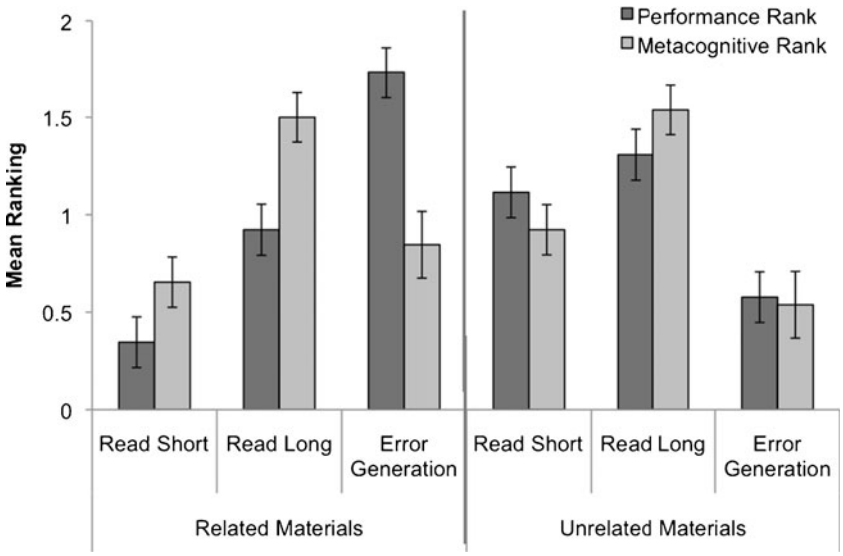


**Fig. 2** Metacognitive Data for Experiment 1 (between-participants). Mean ranking of Learning conditions based on correct performance on the final cued recall test and subjective metacognitive judgments. The condition with the highest proportion correct or subjectively rated the best was assigned a score of a 2. Second best was assigned a score of 1 and worst was assigned a 0

## Discussion

First, consistent with Kornell et al.'s (2009) study, we showed that producing an error for semantically related materials led to enhanced retention. We also found that error-generation did not enhance recall if the materials were completely unrelated. The semantic relation between the cue and target appeared to be critical in determining whether error-generation enhanced memory or not.

A question one might ask is whether participants were behaving similarly when they generated their errors and responded to the feedback in the related and unrelated materials conditions. Perhaps participants were simply guessing randomly and were not sufficiently engaged in the unrelated materials condition, while they were employing all of their efforts to try to generate the answers in the related materials condition. An attentional explanation has been proposed in other error correction paradigms (Butterfield & Mangels, 2003 Butterfield & Metcalfe, 2006; Fazio & Marsh, 2009). Izawa (1967, 1970) has specifically argued that previous errors led to increased learning because of enhanced attention to the corrective feedback. Motivational/attentional differences between conditions might be revealed by the nature of their guesses. By examining the nature of the error responses that the participants produced, we could potentially gain some insight into whether participants' behavior was substantively different behavior was when they generated their errors in the related and unrelated materials conditions.

*Latent semantic analysis* We obtained estimates of the relation between the cues and the generated errors by using latent semantic analysis (LSA). LSA (see Landauer, Foltz & Laham, 1998) is a method of extracting the contextual-usage meaning of words by statistical computations applied to a large corpus of text (Landauer & Dumais, 1997). The aggregate appearance of all words provides a set of mutual constraints that is thought to determine the similarity of meaning of words to one another, given as a cosine. Using LSA (through http://cwl-projects.cogsci.rpi.edu/msr/; see Veksler, Grintsvayg, Lindsey & Gray 2007), it was found, as expected, that the mean relatedness between the cues and targets was higher for related materials ($M$ = .27, $SD$ = .04), than for unrelated materials ($M$ = .05, $SD$ = .01), $t(58)$ = 30.46, $SE$ = .01, $p < .001$. Of more interest, we used LSA to investigate the association between the cue and the error that was generated, in the related and unrelated materials conditions. As is shown in Table 2, when presented with the cue, participants produced errors that were related to the cue in both the related and the unrelated materials conditions. The mean relatedness between cue and generated error for related materials ($M$ = .28, $SD$ = .09) was numerically only slightly higher than that for unrelated materials ($M$ = .25, $SD$ = .09), $t(58)$ = 1.92, $SE$ = .02, $p$ = .056). To see whether participants in the unrelated materials condition altered their guessing strategy as the experiment progressed, the mean association values for the first 15 items were compared with those for the last 15 items. There was no difference in the LSA values for the later trials ($M$ = .24, $SD$ = .09) from those for the earlier trials ($M$ = .26, $SD$ = .08), $t$ = 1.09.

**Table 2** Latent Semantic Analysis (LSA, a semantic relation tool) enabled analysis of the semantic relatedness between the Errors produced by the participant to the provided Cues and Targets. Mean cosine values (the measure provided by LSA) between word pair comparisons are presented below. Higher values indicate a higher degree of semantic relation. These data are presented as a function of Materials condition and Accuracy on the final cued recall test. Standard deviations are provided in the parentheses

| | Cue to Error | | Target to Error | |
|---|---|---|---|---|
| Materials | Correct | Incorrect | Correct | Incorrect |
| *Experiment 1* | | | | |
| Related | 0.30 (.07) | 0.25 (.10) | 0.20 (.04) | 0.19 (.10) |
| Unrelated | 0.25 (.11) | 0.25 (.08) | 0.09 (.04) | 0.07 (.02) |
| *Experiment 2* | | | | |
| Related | 0.27 (.09) | 0.27 (.07) | 0.20 (.13) | 0.22 (.06) |
| Unrelated | 0.34 (.17) | 0.32 (.06) | 0.06 (.05) | 0.06 (.02) |

As was noted in the introduction, we hypothesized that the relation between the generated error and the target might be a critical factor in determining whether error-generation would be beneficial for memory—a possibility that we could also investigate using LSA. Table 2 shows the mean association values for the target–error relation as a function of materials. And indeed, as was hypothesized, the error was more related to the target in the related materials condition ($M$ = .20, $SD$ = .04) than in the unrelated materials condition ($M$ = .08, $SD$ = .02), $t(58)$ = 17.16, $SE$ = .01, $p < .001$.

*Metacognitive illusion* The metacognitive results were particularly interesting. These retrospective judgments were taken after the participants had already had considerable experience with the task. Although participants had just completed the final test moments earlier, those participants in the related materials condition did not realize that the error-generation condition led to the best performance. Instead, they erroneously thought that the read-long condition was the most beneficial for memory of the target items—failing, rather dramatically, to appreciate the benefits of making errors. Furthermore, although performance in each of the three different learning conditions varied greatly between materials, the metacognitive ratings were similar. Comparing materials conditions, what is clear is that although the performance follows two distinct patterns,

the metacognitive ratings do not vary as a function of material relatedness. The metacognitive rankings for each learning condition (read-short, read-long, and error-generation) revealed no statistical differences across materials, ($z$s < 1.60, $p$s > .13). Therefore, although we see a performance boost from error-generation for related materials, participants' rankings are no different from those in the unrelated condition. This metacognitive illusion, it seems, is stable and unaffected by the participant's own contradictory experience with the results of the learning task.

## Experiment 2

In the second experiment, we endeavored to replicate the results of Experiment 1 in a within-participants design, to address more fully the question of why there was a benefit of error-generation only when the cue and target were semantically related. One motivation for a within-participants design was that randomly mixing the presentation of related and unrelated materials would ensure that participants were cognitively engaging in similar tasks when generating an error and would obviate the small difference in response to the cues seen in the LSA analysis in Experiment 1. In the within-participants design, when only the cue was displayed on the screen, the participants could not know whether the forthcoming target would be related or unrelated to the cue. If the lack of memorial benefit for unrelated materials from error-generation was an artifact only of overall lack of engagement or attention, a benefit of generating errors might occur for both related and unrelated materials in the within-participants design. Only after error-generation could participants know the relation of the cue and the target. Conversely, if we replicated the results seen in Experiment 1, this would provide stronger evidence that the semantic relation between the error and the target is central in determining when error-generation helps memory.

### Method

*Participants* Thirty-six Columbia students participated for credit. Six nonnative English speakers were excluded, leaving 30 participants' data. Mean age was 20.7 years ($SD$ = 3.3), and 50% of the participants were females.

*Design and materials* A 2(materials: related, unrelated) × 3 (learning: read-short, read-long, error-generation) within-participants design was used. Forty-five of the related material items and 45 of the unrelated material items from Experiment 1 were randomly selected for use in the present experiment, for a total of 90 word pairs. For both related and unrelated materials, three sets of 15 word pairs were created and counterbalanced over participants so that each word pair was assigned to each of the three learning conditions equally.

*Procedure* The procedure was the same as that in the previous experiment. During the study phase, item presentation order was randomized, and, as noted above, items were preassigned to conditions, which were counterbalanced between participants. Order of item presentation was also randomized on the final cued recall test. All instructions were identical to those given in Experiment 1, with the exception of the metacognitive ratings. Since the present design had six conditions, all six were described in the instructions before the participants ranked them in order of best final test performance to worst.

### Results

*Learning phase performance* Participants did not correctly answer any of the unrelated materials in the error-generation condition during the learning phase. They correctly guessed 3% of the related targets ($SD$ = .03). All results from the error-generation condition excluded the trials for which participants guessed correctly on the initial test.

*Final cued recall test performance* As is shown in Fig. 1, there was an interaction between learning condition and materials. Error-generation led to the highest correct performance for related materials, but it did not lead to benefits with unrelated materials, $F(2, 58) = 7.89$, $MSE$ = .01, $p < .01$, $\eta_p^2 = .92$. Pairwise comparisons showed that error-generation for related materials led to higher correct recall ($M$ = .70, $SD$ = .20) than did both read-short, $t(29)$ = 4.08, $SE$ = .04, $p < .001$, and read-long, $t(29) = 2.51$, $SE$ = .04, $p < .05$, which did not differ from one another, $t(29)$ = 1.61, $SE$ = .04, $p = .12$. There were no significant pairwise differences in performance for the three learning conditions with unrelated materials (all $t$s < 1). As was expected, participants remembered more of the correct targets for the related than for the unrelated materials, $F(1, 29) = 344.32$, $MSE$ = .03, $p < .001$, $\eta_p^2 = .21$. Although qualified by the interaction, there was a main effect of learning condition such that error-generation led to the highest correct performance overall, followed by read-long and read-short, $F(2, 116) = 4.06$, $MSE$ = .03, $p < .05$, $\eta_p^2 = .12$.

*Reaction times* Table 1 shows mean RTs as a function of accuracy on the final cued recall test, learning and material conditions. Only 16 participants had observations for all cells. Overall, items answered correctly ($M$ = 3.69 s, $SD$ =

1.34) were produced more quickly than incorrect items ($M$ = 7.04 s, $SD$ = 2.84), $F(1, 29) = 69.18$, $MSE$ = 14.59, $p < .001$, $\eta_p^2 = .71$. When participants previously made an incorrect guess in the error-generation condition ($M$ = 6.01 s, $SD$ = 2.25), their subsequent RTs on the final cued recall test were longer than those in the read-short ($M$ = 5.02 s, $SD$ = 2.19) and the read-long ($M$ = 5.08 s, $SD$ = 1.82) conditions, $F(2, 58) = 6.88$, $MSE$ = 5.31, $p < .01$, $\eta_p^2 = .19$. Items in the error-generation condition that were answered incorrectly on the final test took longer to produce than items answered correctly, $F(2, 58) = 5.73$, $MSE$ = 4.35, $p < .01$, $\eta_p^2 = .17$. The relatedness of the materials did not lead to differing response latencies on the final test, $F < 1$, nor did materials interact with any other factor.

*Error persistence* For unrelated materials, 15% of the responses on the cued recall test were original errors that persisted from the learning phase to the final test. The original errors persisted only 7% of the time for related materials, $t(29) = 3.65$, $MSE$ = .01, $p < .01$.

*Metacognition* Performance for each condition was ranked from best to worst. Because there were six conditions in the experiment, the best condition for each participant was assigned a score of 5, and the worst was assigned 0. As can be seen in Fig. 3, the error-generation condition for related materials was objectively the best condition for retention ($M$ = 4.38, $SD$ = .87). However, this condition was given only a mean metacognitive ranking of 2.53 ($SD$ = 1.54), $z = 4.12$, $p < .001$. Conversely, although read-long for related materials was subjectively believed to have produced the best performance ($M$ = 4.53, $SD$ = 1.03), in fact, it most often led to worse performance than did error-generation ($M$ = 3.72, $SD$ = 1.07). Noticeably, the metacognitive ranking and performance ranking for read-long are not aligned, $z = 3.34$, $p < .001$. Finally, mean metacognitive judgments indicated that participants subjectively believed that the unrelated read-long condition led to better performance than it actually did ($M_{metacognitive}$ = 2.37 $SD$ = 0.82; $M_{performance}$ = 1.37 $SD$ = 0.97), $z = 3.48$, $p < .01$. No significant differences were found between the mean metacognitive and performance rankings for the three other cells (related–read-short, unrelated–read-short, and unrelated–error-generation), $z$s < 1.

*Latent semantic analysis* The results from the LSA mirror those of Experiment 1, despite many participants being excluded from the analysis due to lack of observations in every cell (see Table 2). Participants generated errors that were related to the cue regardless of materials condition. The mean relation value between the cue and error was slightly higher for unrelated materials ($M$ = .32, $SD$ = .06) than for related materials, ($M$ = .26, $SD$ = .05) $t(29) = 5.06$, $SE$ = .01, $p < 001$, although at the time of generating the error, the participant could not be aware of the subsequent relation to the target, and this effect is in the opposite direction in Experiment 1. The semantic relatedness of the errors to the cues provided support for the idea that participants were engaged and truly generating reasonable errors, even for the unrelated materials. As was expected, errors in the unrelated materials condition were not as related to the target ($M$ = .06, $SD$ = .02) as were errors generated for related materials ($M$ = .21, $SD$ = .06), $t(29) = 17.51$, $SE$ = .01, $p < .001$.

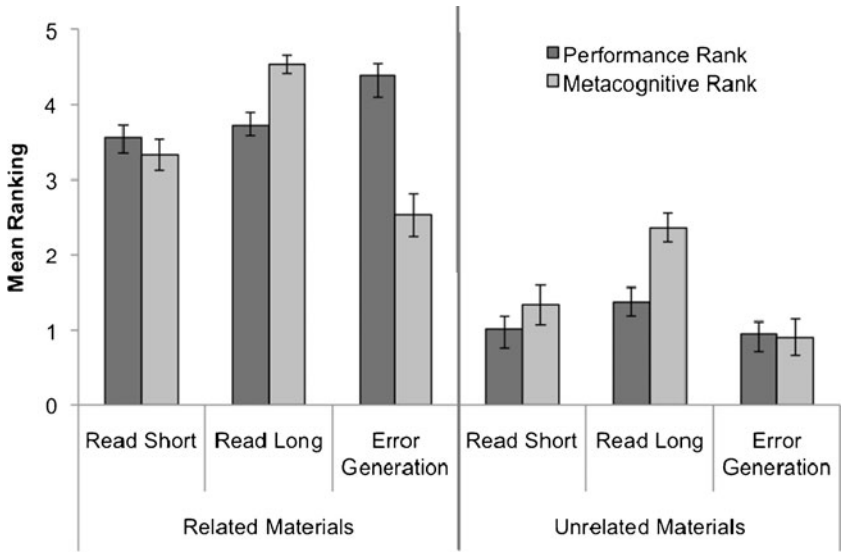


**Fig. 3** Metacognitive Data for Experiment 2 (within-participants). Mean ranking of Learning conditions based on correct performance on the final cued recall test and subjective metacognitive judgments. The condition with the highest proportion correct or subjectively rated the best was assigned a score of a 5. Second best was assigned a score of 4 and so forth, while the worst condition was assigned a 0

## Discussion

Experiment 2 replicated the findings of Experiment 1: Error-generation led to memorial benefits over both reading conditions, but only for related materials. For semantically related word pairs in both experiments, there was enhanced retention for the correct response when participants had made a prior incorrect response, as compared with when they had just read the word pairs. For the unrelated materials, there was no such benefit of incorrect guessing in either experiment. If the benefit from producing an error was due to the effort or engagement during generation itself, there should have been some benefit from incorrect guessing for the unrelated materials. Insofar as items were randomized, in the second experiment, participants could not have been aware of what the next trial would be. Therefore, the processing was the same across related and unrelated conditions during the act of generating the error itself. The results from the LSA substantiated this lack of difference during error-generation. Therefore, it appears that the differential benefits of error-generation between related and unrelated materials began at the time of target feedback.

## General discussion

These results support the idea that semantic closeness is a critical factor in determining whether an error will or will not help learning. One framework consistent with these results is the Osgood (1949) transfer surface, which captured all transfer of learning relations that were known at the time of publication. In this surface, similarity between intralist stimuli (cues) is plotted against similarity of intralist responses (targets). Of importance is how these two factors interact to produce positive transfer or, conversely, interference. When cues are identical, as in our experiments, the more related the responses are to one another, the greater the positive transfer will be. Thus, this framework would predict that positive transfer will result if the error and target are related. It is only when the two responses—the error and the target—are unrelated that negative transfer or interference should be produced. For the error-generation condition for related materials in our experiments, LSA showed that errors were highly similar to the correct targets. Therefore, Kornell et al.'s (2009) materials and our related materials condition conformed to Osgood's (1949) A–B A–B' situation. The erroneous answer produced in this context facilitated learning of the correct answer, B'. Conversely, the LSA ratings showed that our unrelated materials condition conformed to Osgood's A–B, A–C situation. The errors, in that condition, were unrelated to the correct targets and produced no memory benefit. In fact, there was a slight suggestion of error-related interference. As compared with related materials, unrelated materials led to more of the original errors persisting in the final test. Additionally, in the first experiment, correct item recall was worse in the error-generation condition than in either the read-long or read-short condition.

Since the time of Osgood (1949), two possible explanations have emerged for why this relationship between the error and the target might be important. One explanation is that making a related error helps form a richer, more elaborate network with the cue and the error, as compared with an unrelated error. In terms of levels of processing, encoding in a deeper, more elaborative manner is beneficial for future retrieval (Craik & Lockhart, 1972; Craik & Tulving, 1975). Through elaborative processing, by producing a guess and forming an elaboration based on a "deep" or semantic level, retention is enhanced above "shallow" processing. Error-generation of a related item might be an elaboration, making the target more meaningful. Although one might engage in elaborative processes for unrelated materials, this elaboration might be in vain. For example, when provided with the word *attack*, when one tries to generate a response, one will presumably think about what it means, and generate, erroneously, *dog*. When the related target, *defend*, is displayed, the connection is clear, and one can draw a more elaborate and meaningful relationship than when one simply sees *attack–defend*. One can imagine an attack dog defending his doghouse, defending oneself against an attacking dog, or both. This richer, more elaborate encoding method should help retention. However, if the correct answer is something unrelated to attack, such as *bicycle*, it is more difficult to form a meaningful connection or elaboration between the cue, error, and target. Additionally, Carpenter (2009) and Carpenter and DeLosh (2006) have argued that elaboration is less likely to occur when one is reading, as compared with active retrieval.

Along similar lines, during error-generation, one might activate a variety of related concepts that provide a more elaborate, richer memory trace, consistent with spreading activation theories of memory (e.g., Collins & Quillian, 1972). Since there is more information that could potentially activate the correct target, this elaborative structure could aid recall (e.g., J. R. Anderson, 1983). Carpenter (2011) suggests that retrieval aids in activating semantically related information above restudy. In other recent work, Grimadli and Karpicke (in press) found error-generation benefits only for semantically related items, a finding consistent with the results presented in the present article. Conversely, when participants' errors were constricted by providing the first few letters of the error (e.g., *tide–wa____*), an error-generation benefit was not obtained. The authors interpret their results as favoring a spreading activation view—that is, when an error is committed, concepts that are related to the target are activated and enhance learning (e.g., Collins & Loftus, 1975).

A mediator hypothesis is a second potential explanation for why the relation between the error and the target may be

important in determining whether errors benefit recall of the correct target. Under some circumstances, the error itself may serve as a mediator, or secondary link, between the cue and the target. It has been shown that previous retrieval attempts can serve as an intermediary cue in target retrieval (Soraci et al., 1999) and can facilitate recall (Pyc & Rawson, 2010). The latter authors found beneficial effects of the mediator, however, only when it both could be retrieved at time of test and elicited the target item.[3] In the present paradigm, it seems more likely that a word that is related to the target might serve as an effective mediator than would one that is unrelated to the target.

These two hypotheses—*error as an elaboration* and *error as a mediator*—are not mutually exclusive. *Error as an elaboration* suggests that because of enhanced processing at encoding from an active (elaborative processing) or passive (semantic activation) process, the correct target will be remembered better when an error is generated than with simple study. In addition, even at retrieval, those concepts that were previously activated might lead to enhanced recall of the correct target. On the other hand, the *error as a mediator* hypothesis suggests that recalling the original error itself, and not just the surrounding semantic landscape, can act as a secondary cue to retrieve the target.

The RT data are readily interpretable within the *error as a mediator* hypothesis. Participants took longer to produce a response on the final test for the error-generation condition, as compared with the read-long and read-short conditions. When they attempted to retrieve the correct target, the incorrect guesses might have served as a secondary link that introduced a second step into the retrieval process. This second step would require additional time, thereby leading to longer RTs. Even for unrelated materials, if one retrieved the original error and tried to use it as a mediator, the response time would still be longer due to the additional, unsuccessful labor involved in trying to find the correct retrieval path to the target.

The RT data could also be interpreted within the *error as an elaboration* view, though, insofar as exploring the elaborations that were set up at encoding could be assumed to take time. A number of semantic activation models predict longer RTs with a higher number of associated concepts (see ACT–R and the fan effect; J. R. Anderson, 1974; J. R. Anderson & Reder, 1999). These models could also predict that participants' RTs would be longer for the error-generation conditions as a result of response competition between the original generated error and the correct target.

Finally, in both experiments, the metacognitive data show a stable illusion, whereby participants were not aware that error-generation was helpful for remembering related word pairs. It is, perhaps, not surprising that committing errors during learning is typically seen in a negative light. As Bjork (1994) stated, "Errors made during training are generally not viewed as opportunities for learning, but rather, as evidence of a less-than-optimal training program" (p. 299). It is surprising, however, that even moments after completion of the criterion test, participants were not aware that error-generation was beneficial for related materials. This finding is particularly interesting since these global retrospective judgments of performance can use information acquired during the criterion test to help inform judgments. Retrospective judgments, therefore, have been shown to be more accurate than predictions of performance (see Pieschl, 2009). For this reason, it is particularly interesting that there seems to be such a large disconnect between subjective performance rankings and actual performance.[4]

Although, currently, we cannot make any claims in regard to potential mechanisms driving the subjective bias against errors, this bias is still of great interest. There are several possible explanations for the error-generation metacognitive illusion. One is that participants simply had a bias against believing that errors are beneficial. A second explanation is that participants relied on a *ease of processing* heuristic (see Koriat & Ma'ayan, 2005; Winkielman, Schwarz, Fazendeiro & Reber, 2003) or, more specifically, *easily learned, easily remembered* (Koriat, 2008; Miele & Molden, 2010). There have been a number of experiments in which how easily stimuli are processed influences judgments of how well information is learned (e.g., Carpenter & Olson in press; Koriat, 1997, 2008; Nelson & Dunlosky, 1991; Rawson & Dunlosky, 2002; Rhodes & Castel, 2008). Since error-generation might not have seemed as easy (perhaps both at retrieval and at encoding) as reading the answer, participants might be underconfident in this strategy. Furthermore, if participants were also generating the error as a mediator, despite its beneficial effect on retention, the presence of another potential competitor could have driven down performance estimates.

From an educational standpoint, the findings of the two reported experiments are of relevance for two reasons. First, we have shown that when the materials are related, even when that relation is very small—low associates, not high associates—

[3] Although more original errors were produced on the final test for unrelated materials in Experiment 1, this does not necessarily mean that those in the related materials condition were not capable of retrieving their original error. Anecdotally, during the debriefing, many participants in the related materials condition mentioned that they remembered their guesses.

[4] It is possible, from the present data, that participants underestimated the memorial benefits of generating errors because they could not remember which items were in the error-generation condition (although cf. Huelser & Metcalfe, 2011).

generating an error and receiving corrective feedback is much better for learning than is simply studying. Although more research must be done to understand the exact mechanisms behind the error-generation effect, the present results suggest that guessing should be encouraged, even if the result is an error. Rarely will the question and answer be so far removed that the learner cannot make a meaningful connection between the two. However, some caution is needed in implementing this recommendation, given that errors may have detrimental effects for memory-impaired individuals, as was shown in Clare and Jones's (2008) review. It is not yet known whether error-generation, when the errors are related to the targets, as in the present study, will lead to enhanced or diminished performance for young children or those with learning disabilities.

The second point of interest to educators comes from the metacognitive monitoring results. It is clear that even immediately after completion of the criterion test, participants were not aware of which study strategy was best for learning. It is quite plausible that learners rely on these types of global retrospective judgments when deciding what learning strategy to use. It has been shown that monitoring has consequences for metacognitive control, or regulation, of learning (Metcalfe & Finn, 2008; Son, 2004; Son & Kornell, 2008; Son & Metcalfe, 2000; Stone, 2000; see also Metcalfe, 2009). Thus, it seems unlikely that the learner, without further training of his or her metacognition, will implement this highly effective learning strategy.

**Author note** This research was supported by Grant 220020166 from the James S. McDonnell Foundation, and by a National Science Foundation graduate fellowship to the first author. We thank Brandon Luke and Margaret Lee for assistance with material creation and data collection. We also thank Lisa Son, Karen Kelly, David Miele, Peter Messa, Joseph Bisoglio, Katherine Rawson, Jeff Karpicke, Sean Kang, and an anonymous reviewer for their helpful comments.

## References

Anderson, J. A. (1973). A theory for the recognition of items from short memorized lists. *Psychological Review, 80,* 417–438.

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology, 6,* 451–474.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Behavior, 22,* 261–295.

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review, 79,* 97–123.

Anderson, M. C., & Neely, J. H. (1996). Interference and inhibition in memory retrieval. In E. L. Bjork & R. A. Bjork (Eds.), *Memory: Handbook of perception and cognition* (2nd ed., pp. 237–313). San Diego, CA: Academic Press.

Anderson, J. R., & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General, 128,* 186–197.

Barnes, J. M., & Underwood, B. J. (1959). "Fate" of first-list associations in transfer theory. *Journal of Experimental Psychology, 58,* 97–105.

Berger, S. A., Hall, L. K., & Bahrick, H. P. (1999). Stabilizing access to marginal and submarginal knowledge. *Journal of Experimental Psychology: Applied, 5,* 438–447.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A metanalytic review. *Memory & Cognition, 35,* 201–210.

Bjork, R. A. (1994). Institutional impediments to effective training. In D. Druckman & R. A. Bjork (Eds.), *Learning, remembering, believing: Enhancing human performance* (pp. 295–306). Washington, DC: National Academy Press.

Butler, A. C., Karpicke, J. D., & III Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 918–928.

Butterfield, B., & Mangels, J. A. (2003). Neural correlates of error detection and correction in a semantic retrieval task. *Cognitive Brain Research, 17,* 793–817.

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition and Learning, 1,* 69–84.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1563–1569.

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(6), 1547–1552

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34,* 268–276.

Carpenter, S. K., & Olson, K. M. (in press). Are pictures good for learning new vocabulary in a foreign language? Only if you think they are not. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Clare, L., & Jones, R. S. P. (2008). Errorless learning in the rehabilitation of memory impairment: A critical review. *Neuropsychology Review, 1,* 1–23.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*, 407–428.

Collins, A. M., & Quillian, M. R. (1972). Experiments on semantic memory and language comprehension. In L. Gregg (Ed.), *Cognition and learning* (pp. 117–138)). New York: Wiley.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671–684.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 10,* 268–294.

Cunningham, D., & Anderson, R. C. (1968). Effects of practice time within prompting and confirmation presentation procedures on paired associate learning. *Journal of Verbal Learning and Verbal Behavior, 7,* 613–616.

Eich, J. M. (1982). A composite holographic associative recall model. *Psychological Review, 89,* 627–661.

Fazio, L. K., Huelser, B. J., Johnson, A., & Marsh, E. J. (2010). Receiving right/wrong feedback: Consequences for learning. *Memory, 18,* 335–350.

Fazio, L. K., & Marsh, E. J. (2009). Surprising feedback improves later memory. *Psychonomic Bulletin & Review, 16,* 88–92.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1–67.

Glaser, R. (1990). The reemergence of learning theory within instructional research. *American Psychologist, 45,* 29–39.

Grimadli, P. J. & Karpicke, J. D. (in press). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*.

Guthrie, E. (1952). *The psychology of learning* (Revth ed.). New York: Harper.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers, 16,* 96–101.

Huelser, B. J. & Metcalfe, J. (2011, November). Performance monitoring offsets, but does not eliminate, the metacognitive illusion that errors hurt learning. *Poster presented at the 52nd annual meeting of the Psychonomic Society, Seattle, WA*.

Izawa, C. (1967). Function of test trials in paired-associate learning. *Journal of Experimental Psychology, 75,* 194–209.

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology, 83,* 340–344.

Jarvis, B. G. (2004). DirectRT (Version 2004.1.0.55) [computer software]. New York: Empirisoft Corporation.

Kane, J. H., & Anderson, R. C. (1978). Depth of processing and interference effects in the learning and remembering of sentences. *Journal of Educational Psychology, 70,* 626–635.

Kang, S. H. K., Pashler, H., Cepeda, N. J., Rohrer, D., Carpenter, S. K., & Mozer, M. C. (2011). Does incorrect guessing impair fact learning? *Journal of Educational Psychology., 131,* 48–59.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126,* 349–370.

Koriat, A. (2008). Easy comes, easy goes? The link between learning and remembering and its exploitation in metacognition. *Memory & Cognition, 36,* 416–428.

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language, 52,* 478–492.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 989–998.

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory, 17,* 493–501.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284.

Loftus, E. F. (1979). *Eyewitness testimony*. Cambridge, MA: Harvard University Press.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L., III. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology, 103,* 399–414.

McGeoch, J. A. (1942). *The psychology of human learning*. New York: Longmans.

Melton, A. W., & Irwin, J. M. (1940). The influence of degree of interpolated learning on retroactive inhibition and the overt transfer of specific responses. *American Journal of Psychology, 53,* 173–203.

Metcalfe, J. (1990). Composite holographic associative recall model (CHARM) and blended memories in eyewitness testimony. *Journal of Experimental Psychology: General, 119,* 145–160.

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science, 18,* 159–163.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15,* 174–179.

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors and feedback. *Psychonomic Bulletin & Review, 14,* 225–229.

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition, 37,* 1077–1087.

Miele, D. B., & Molden, D. C. (2010). Naïve theories of intelligence and the role of processing fluency in perceived comprehension. *Journal of Experimental Psychology: General, 139,* 535–557.

Mulligan, N. W., & Lozito, J. P. (2005). Self-generation and memory. In B. H. Ross (Ed.), *Psychology of learning and motivation* (pp. 175–214). San Diego: Elsevier.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL-effect." *Psychological Science, 5*, 207–213.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms*. Retrieved from http://w3.usf.edu/FreeAssociation/

Osgood, C. E. (1949). The similarity paradox in human learning: A resolution. *Psychological Review, 56,* 132–143.

Parlow, J., & Berlyne, D. E. (1971). The effect of prior guessing on incidental learning of verbal associations. *Journal of Structural Learning, 2,* 55–65.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 3–8.

Pashler, H., Zarow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1051–1057.

Pavio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology, 76,* 1–25.

Pieschl, S. (2009). Metacognitive calibration—an extended conceptualization and potential applications. *Metacognition Learning, 4,* 3–31.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330,* 335.

Rawson, K. A., & Dunlosky, J. (2002). Are performance predictions for text based on ease of processing. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 28,* 69–80.

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137,* 615–625.

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied, 15,* 243–257.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 592–604.

Slamecka, N. J., & Fevreiski, J. (1983). The generation effect when generation fails. *Journal of Verbal Learning and Verbal Behavior, 22,* 153–163.

Son, L. K. (2004). Spacing one's study: Evidence for a metacognitive control strategy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 601–604.

Son, L. K., & Kornell, N. (2008). Research on the allocation of study time: Key studies from 1890 to the present (and beyond). In J. Dunlosky & R. A. Bjork (Eds.), *A handbook of memory and metamemory* (pp. 333–351). Hillsdale, NJ: Psychology Press.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 204–221.

Soraci, S. A., Jr., Carlin, M. T., Chechile, R. A., Franks, J. J., Wills, T., & Watanabe, T. (1999). Encoding variability and cuing in generative processing. *Journal of Memory & Language, 41,* 541–559.

Stone, N. J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review, 12,* 437–475.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95,* 66–73.

Veksler, V. D., Grintsvayg, A., Lindsey, R., & Gray, W. D. (2007). A proxy for all your semantic needs. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Meeting of the Cognitive Science Society* (p. 1878). Austin, TX: Cognitive Science Society. Retrieved from http://cwl-projects.cogsci.rpi.edu/msr/

Webb, L. W. (1917). Transfer of training and retroaction: A comparative study. *Psychological Monographs, 24,* 1–90.

Winkielman, P., Schwarz, N., Fazendeiro, T., & Reber, R. (2003). The hedonic marking of processing fluency: Implications for evaluative judgment. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 189–217). Mahwah, NJ: Erlbaum.

Neurobiology of Sleep and Circadian Rhythms 3 (2017) 10–16

Contents lists available at ScienceDirect

Neurobiology of Sleep and Circadian Rhythms

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

Research paper

# Sleep deprivation impairs recognition of specific emotions

William D.S. Killgore[a,b,*], Thomas J. Balkin[b], Angela M. Yarnell[b], Vincent F. Capaldi II[b]

[a] University of Arizona, USA
[b] Walter Reed Army Institute of Research, USA

ARTICLE INFO

*Keywords:*
Sleep deprivation
Emotion recognition
Facial expression
Perception

ABSTRACT

Emotional processing is particularly sensitive to sleep deprivation, but research on the topic has been limited and prior studies have generally evaluated only a circumscribed subset of emotion categories. Here, we evaluated the effects of one night of sleep deprivation and a night of subsequent recovery sleep on the ability to identify the six most widely agreed upon basic emotion categories (happiness, surprise, fear, sadness, disgust, anger). Healthy adults (29 males; 25 females) classified a series of 120 standard facial expressions that were computer morphed with their most highly confusable expression counterparts to create continua of expressions that differed in discriminability between emotion categories (e.g., combining 70% happiness+30% surprise; 90% surprise+10% fear). Accuracy at identifying the dominant emotion for each morph was assessed after a normal night of sleep, again following a night of total sleep deprivation, and finally after a night of recovery sleep. Sleep deprivation was associated with significantly reduced accuracy for identifying the expressions of happiness and sadness in the morphed faces. Gender differences in accuracy were not observed and none of the other emotions showed significant changes as a function of sleep loss. Accuracy returned to baseline after recovery sleep. Findings suggest that sleep deprivation adversely affects the recognition of subtle facial cues of happiness and sadness, the two emotions that are most relevant to highly evolved prosocial interpersonal interactions involving affiliation and empathy, while the recognition of other more primitive survival-oriented emotional face cues may be relatively robust against sleep loss.

## 1. Introduction

There is an emerging consensus that sleep plays a vital role in recalibrating the emotional functioning of the brain (Walker and Van Der Helm, 2009; Walker, 2009). Without sufficient sleep, there appears to be a reduction in emotional regulation capacities and a loss of perceptual sensitivity to cues that provide critical emotional information about the external environment and internal milieu (Goldstein-Piekarski et al., 2015). Notably, Yoo and colleagues showed that compared to the sleep-rested state, sleep deprivation was associated with increased amygdala responses to negatively valenced visual images (e.g., mutilated bodies; unsanitary conditions; aggressive scenes), and reduced functional connectivity between the top down emotion regulating regions of the medial prefrontal cortex and the emotionally responsive amygdala (Yoo et al., 2007). The findings suggest that without sleep, there is a weakening of the ability of higher order brain regions to exert regulatory control over more primitive threat detection systems, leading to greater emotional reactivity. In a parallel study from the same lab, Gujar and colleagues demonstrated that sleep deprivation produced similar increases in limbic and paralimbic regions to positively valenced images as well (Gujar et al., 2011), suggesting that sleep loss increases emotional reactivity to both positive and negative stimuli. This has led to the suggestion that sleep deprivation may globally lower the threshold for emotional activation, regardless of valence, thus increasing overall sensitivity to emotional stimuli (Simon et al., 2015).

The effects of sleep deprivation are not limited to emotionally evocative scenes, but also affect how the human brain responds to facial expressions of emotion. Huck and colleagues conducted one of the earliest investigations of the effects of sleep deprivation and stimulant countermeasures on the ability to accurately identify emotional displays depicted in photographs of facial expressions (Huck et al., 2008). In that study, participants completed two tasks, one involving categorization of simple photographs of six basic emotions, and the other involving categorization of complex blended images created by morphing pairs of highly confusable emotions (e.g., fear+surprise) from the same set of six primary emotions. While accuracy for simple emotion perception was not affected by sleep deprivation or stimulants, the ability to accurately identify the dominant emotion within complex emotional blends was adversely affected by sleep deprivation and was

* Correspondence to: Department of Psychiatry, University of Arizona, PO Box 245002, Tucson, AZ 85724, USA.
*E-mail address:* killgore@psychiatry.arizona.edu (W.D.S. Killgore).

http://dx.doi.org/10.1016/j.nbscr.2017.01.001
Received 19 September 2016; Received in revised form 18 December 2016; Accepted 20 January 2017
Available online 21 January 2017

restored by stimulant medications (Huck et al., 2008). Whereas Huck and colleagues did not examine the effects of sleep loss on accuracy for the six specific emotions, a subsequent study by van der Helm and colleagues focused on recognition of three separate emotions, including happy, sad, and angry facial expressions (Van Der Helm et al., 2010). They presented participants with a set of morphed face photographs that ranged in intensity from neutral to full strength for each prototypical emotion. The authors found that sleep deprivation led to a significant impairment of recognition for angry and happy expressions, but only in the middle range of intensity—an effect that was most prominent in female participants. More recently, Cote and colleagues examined a broader range of facial affects, including happy, sad, angry, and fear and found that sleep deprivation impaired recognition of sadness in simple full face expressions as well as faces morphed to a moderate level of intensity (Cote et al., 2014). Finally, Goldstein-Piekarski and colleagues presented participants with a series of computer generated faces differing on a continuum from safe to highly threatening in appearance and found that sleep deprivation led to a bias toward overestimating the threat in faces, a finding that was associated with altered viscerosensory brain activation (Goldstein-Piekarski et al., 2015). Thus, it is clear that sleep deprivation leads to an impairment in recognition of some aspects of facial affect, particularly when there is some ambiguity in the expressions, but there is no consensus regarding the specific emotions that are most sensitive to these effects.

To provide additional insights into this topic, here we provide a further analysis of the data presented in our previous article (Huck et al., 2008). In that paper, we only reported total recognition accuracy scores that were collapsed across all six expressions. However, given the current interest in the topic, we believe that it would be informative to provide additional unpublished data regarding the effects of sleep deprivation and recovery sleep on accuracy for recognizing the dominant emotional expressions in morphed blends of highly confusable emotions based on the six universal facial expressions, including happiness, sadness, surprise, fear, disgust, and anger. Based on the aforementioned literature, we hypothesized that sleep deprivation would lead to sustained accuracy for ambiguous blends of confusable faces with predominantly high threat relevance (i.e., Anger, Fear, Surprise), whereas morphed expressions with predominantly social/affiliative relevance (i.e., Happiness, Sadness) would be most susceptible to degraded recognition from sleep loss.

## 2. Method

### 2.1. Participants

A total of 54 (29 male; 25 female) healthy young adults (Mean Age=23.5, SD=4.0) volunteered for a larger study of the effects of sleep deprivation and stimulant countermeasures on various aspects of cognitive functioning. While related findings from this dataset have been reported elsewhere (Huck et al., 2008), here we present previously unpublished findings and re-analysis of those data regarding the effects of sleep deprivation on the ability to recognize the six universal basic emotions. All participants underwent a physical examination prior to entry into the study and were deemed to be physically healthy by the examining physician. Exclusionary criteria included any history of sleep disorder, psychiatric illness, drug or alcohol abuse, cardiac problems, current pregnancy, or other health issue that would pose a risk for participating in a sleep deprivation study. Volunteers were also excluded for any history of tobacco use in the preceding three years or daily caffeine intake in excess of 400 mg/day. Participants were required to abstain from alcohol, stimulants, or other psychoactive drugs for 48 h prior to participation. Abstinence from stimulants was verified via urine drug screen at time of entry into the study and every 24 h thereafter. Written informed consent was obtained prior to enrollment and all participants were compensated for their time in the lab and were also offered a performance bonus for demonstrated effort on all study tasks. The protocol for this study was approved by the Walter Reed Army Institute of Research Human Use Review Committee and the U.S. Army Human Subjects Research Review Board. This material has been reviewed by the Walter Reed Army Institute of Research and there is no objection to its presentation and/or publication.

### 2.2. Materials and procedure

As part of the larger study, participants underwent a four night continuous in-residence laboratory study, which consisted of a baseline acclimation night involving 8-h enforced time in bed from 2300 to 0700, a 61-h period of monitored continuous wakefulness during which time participants completed a variety of cognitive and performance tasks, a 12-h enforced recovery sleep opportunity from 2000 to 0800, and a post-recovery day involving additional cognitive and performance tasks. The present analysis is focused primarily on the outcome of the Ekman Hexagon Test (EHT; Thames Valley Test Company, Suffolk, UK), which was administered several times during the course of the study. The EHT is a computer-administered task that required the participant to classify each of a series of displayed facial expression photographs according to one of six different emotion labels (happiness, surprise, fear, sadness, disgust, anger). As depicted in Fig. 1, each face was displayed on the screen with six different emotion labels located below the face. Label order was randomized at each presentation. The participant used a mouse to click on the label that best represented the displayed emotion for each trial. Each facial photograph was displayed for up to five seconds, after which the photograph disappeared but the labels remained until a response was made. There was no time limit for responses. The EHT comprised 150 trials and all photographs were of the same male individual poser.

To increase the emotion processing demands of the task, the face stimuli were previously morphed to produce a continuum of facial blends. Briefly, as described in the EHT manual, each basic emotion photograph was computer morphed with its two most similar appearing and frequently confused counterpart emotions (e.g., fear was morphed with surprise and fear was morphed with sadness to create a continuum of expressions from surprise to fear to sadness). For example, two face images would be combined to create a new image depicting 70% fear and 30% surprise. Each morphed image was created in by combining two prototype images according to the following ratios: 90:10; 70:30; 50:50; 30:70; 10:90. As depicted in Fig. 2, this process yielded 30 combinations of facial expressions (6 emotions×5 blend levels), that comprised a continuum of emotional blends, with

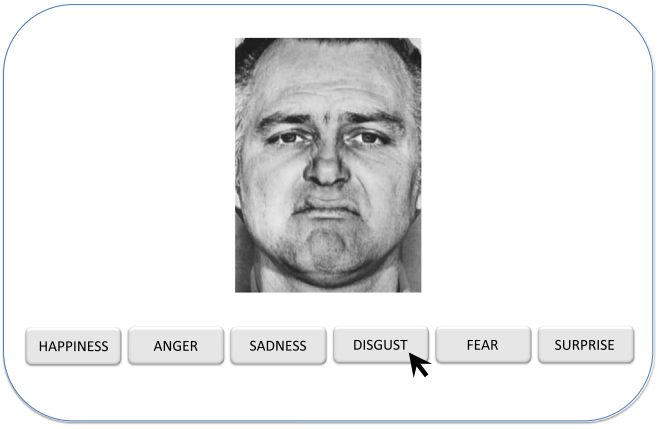


**Fig. 1.** Example of the Emotion Hexagon Test. Participants were shown a series of 150 facial expressions that comprised morphed blends of pairs of highly confusable emotions (e.g., 70% disgust+30% sadness) for five seconds and used the computer cursor to select the most accurate label for each expression.

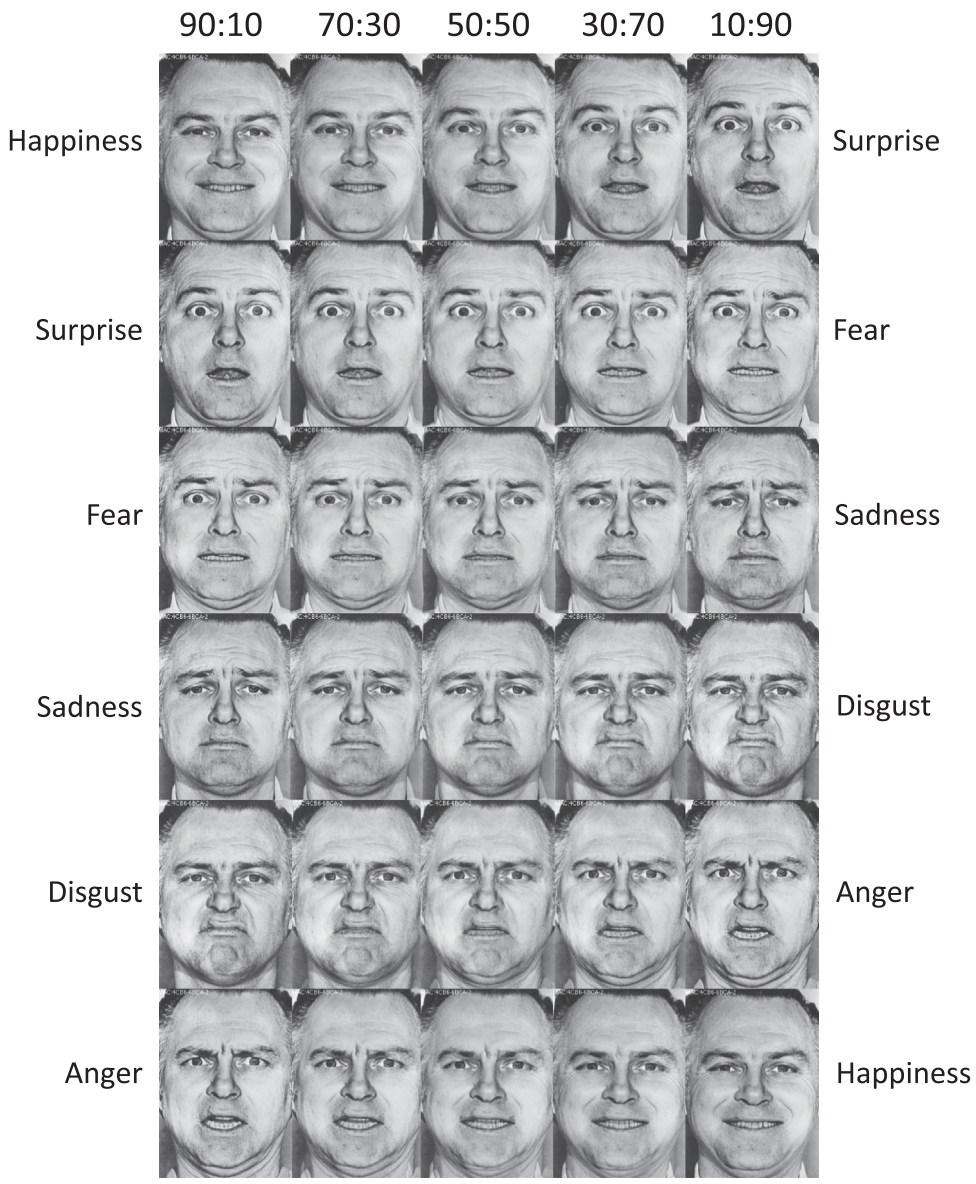


**Fig. 2.** The continuum of faces used for the Emotion Hexagon Test. A total of 30 morphed faces of the same male poser were used. Each face was a morph between two neighboring faces at varying ratios (e.g., 90% happiness+10% surprise). The blended expressions are listed along the sides of the continua and the proportions of each expression included in the face are listed along the top axis.

those closest to the 50:50 ratio being most difficult to discriminate. Although the 50:50 images were presented during the EHT, they were not calculated in the final scores. Since there is no objectively correct response for the 50:50 faces (i.e., either response is correct), those stimuli will not be discussed further here. Thus 120 blended images were presented (20 for each emotion) and recorded as correct or incorrect for each response.

The EHT was administered at 1230 on the baseline day (i.e., after 5.5 h of wakefulness), at 0630 on the first morning following one night of sleep deprivation (i.e., after 23.5 h of wakefulness), again at 0540 on the second morning of sleep deprivation (i.e., after 46.7 h of wakefulness and stimulant medication administration), and finally at 1200 on the final day, following 12 h of recovery sleep (i.e., after 4 h of wakefulness). Data from the 46.7-h post-stimulant session were reported in great detail in a previous publication (Huck et al., 2008) and will not be further analyzed or discussed in the present article. Thus, we report here data for baseline, one night of sleep deprivation (no stimulants), and recovery.

### 2.3. Analysis

Data were analyzed in IBM SPSS 20. The raw number of correct items was re-calculated in terms of percent correct for each category and entered into a 3 (session)×6 (emotion category)×2 (sex) repeated measures analysis of variance (ANOVA). Because the primary effect of interest was whether sleep deprivation would lead to a significant change in emotion recognition accuracy, planned comparisons examined the effect of sleep deprivation relative to the pre-sleep deprivation and post-recovery performances within each of the specific emotion categories using polynomial planned contrasts (i.e., predicting a quadratic effect of reduced performance during SD compared to baseline and recovery). For those showing a significant quadratic effect, planned simple effects paired comparisons between the SD and pre- and post-recovery means were conducted. Reaction time (RT) data from the PVT were converted to a metric of psychomotor speed (i.e., 1/RT×1000). All significance tests were evaluated at $\alpha$=.05.

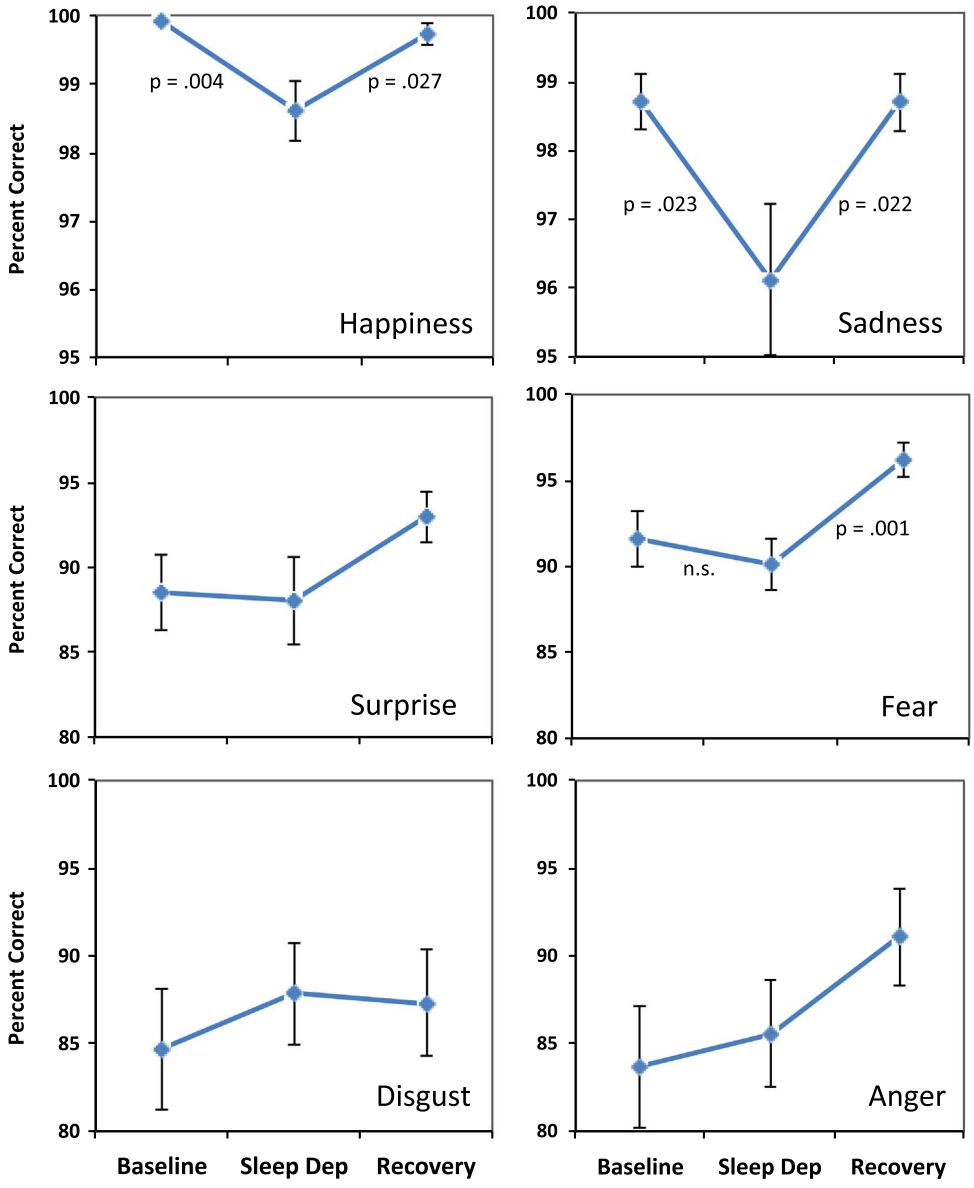


**Fig. 3.** The figures show the accuracy of recognition performance for each blended emotion at baseline, 23.5-h of sleep deprivation, and again following a 12-h opportunity for recovery sleep. The top panels show that sleep deprivation was associated with a significant decline in the percent of correct responses for faces with dominant expressions of happiness or sadness. None of the other emotional expressions showed significant declines in perception accuracy with sleep deprivation.

## 3. Results

The multivariate tests for the repeated measures ANOVA revealed a significant main effect of session, $F(2,51)=8.33$, $p=.001$, a significant main effect of emotion type, $F(5,48)=23.15$, $p < .0001$, and a significant session×emotion type interaction, $F(10,43)=3.69$, $p=001$. Because these findings were not significantly affected by the sex of the participant, as indicated by a non-significant sex×session×emotion type interaction, $F(10,43)=0.83$, $p=.60$, data were combined for the sample as a whole for subsequent analyses.

Fig. 3 shows the plots for the individual emotion categories separately. Planned comparisons showed that for happiness, there was a significant quadratic effect, $F(1,52)=7.12$, $p=.01$, suggesting an effect of sleep deprivation on accuracy judgments. This was confirmed by planned simple effects comparisons showing a decrease in accuracy from baseline to sleep deprivation ($p=.004$) and an increase from sleep deprivation to post-recovery ($p=.027$). Similarly, for sadness, there was a significant quadratic effect, $F(1,52)=5.85$, $p=.019$, and further planned comparisons suggested that this was indeed due to a significant decline in accuracy from baseline to sleep deprivation ($p=.023$) and a significant increase from SD to post-recovery ($p=.022$). On the other hand, for surprise, there was no significant quadratic effect, $F(1,52)=1.73$, $p=.19$, suggesting no effect of sleep deprivation. While there was a significant quadratic effect for fear, $F(1,52)=5.92$, $p=.018$, there was no significant decline from baseline to sleep deprivation ($p=.39$), although there was an increase from sleep deprivation to post-recovery ($p=.001$). No significant quadratic effect was observed for either disgust, $F(1,52)=0.98$, $p=.33$, or anger, $F(1,52)=.68$, $p=.41$, suggesting no effect of sleep deprivation on the accuracy of these emotional judgments.

## 4. Discussion

Human survival has long depended on the ability to accurately read and infer the emotional states of others. Because the face communicates the physical and affective condition, motivation, and potential intentions of those in close proximity, it provides a crucial source of information about factors that could affect survival and wellbeing,

including the presence or absence of danger, the availability of resources, the needs of others, and the potential for social inclusion and affiliation (Blair, 2003). We examined the effect of one night of sleep deprivation on the ability to accurately identify the six most broadly accepted basic human facial expressions of emotion (Ekman, 1992) under conditions of varying ambiguity (i.e., each target emotional expression was partially blended with another emotional expression with which it is often confused). Consistent with our hypothesis, we found that sleep deprivation adversely affected the recognition of target expressions involving happiness and sadness, emotions that often relate most strongly to social and affiliative behaviors, while having no discernable effect on the ability to recognize target facial expressions communicating the potential for immediate threat, peril, or danger (i.e., expressions of surprise, fear, disgust, and anger). Together, these findings suggest that during periods of compromised cognitive-emotional capacity induced by sleep loss, the brain may preserve emotional recognition resources necessary for responding to threat-relevant stimuli at the expense of cognitive resources available for sustaining less urgent socio-emotional recognition processes that play a role in empathy, social closeness, and affiliative behavior.

Our findings contribute to a rapidly emerging literature on the effects of sleep deprivation on emotional processing. Notably, a small number of recent studies have examined the effects of sleep deprivation specifically on the accuracy and perceived intensity of emotional face perception. Evidence suggests that sleep deprivation adversely affects the perception of facial emotion, particularly when facial cues are somewhat ambiguous due to morphing or blending of features (Huck et al., 2008). Most studies that have found effects of sleep deprivation on emotion perception have utilized some sort of computer morphing technique to scale images to varying levels of ambiguity or intensity. Applying such techniques, van der Helm and colleagues found that a night of sleep deprivation reduced accuracy for recognizing angry and happy faces, but only in the middle ranges of intensity and only for females (Van Der Helm et al., 2010). A similar study by Cote and colleagues found that sleep deprivation reduced classification accuracy only for sad faces, particularly when the faces were more ambiguous due to morphing (Cote et al., 2014). Our findings are partially consistent with each of these studies, as we found that sleep loss affected recognition of happy faces (as found by van der Helm and colleagues) and sad faces (as found by Cote and colleagues), while other blended affects were recognized at baseline levels. While it is clear that sleep deprivation affects recognition of emotion, it is not clear why the specific emotions that were affected differed between studies. Notably, our study focused on accuracy data, while the others focused more specifically on neurophysiological methods. Our sample size was larger and included more emotion categories than either of the two preceding studies, which could account for some of the differences. The stimuli for the three studies also differ in terms of the faces used and degree of morphing employed. However, the most likely explanation is that the previous two studies focused on morphed expressions designed to differ on the dimension of intensity (i.e., from neutral to strongly emotional) *within* each emotion category, whereas our approach used expressions that were more ambiguous because they were morphed at full intensity *across* emotion categories with their nearest confusable expressions. Prior work suggests that ambiguity in expressions may activate threat-detection systems (Blasi et al., 2009; Cote et al., 2014), which may explain why the impairments were restricted to only non-threat (i.e., social-affiliative) emotions.

It should also be noted that although sleep deprivation significantly impaired the perception of happy faces, the absolute magnitude of the decline was quite small and there was an overall ceiling effect for perception of happy faces. This is not entirely surprising, as happy faces are well known to be the easiest and most rapidly identified of all expressions (Alves et al., 2009; Wells et al., 2016). This is partially due to the fact that there are more negative emotional categories to choose from than positive ones. In our study, we had five negative emotions and one positive one, which likely made it fairly easy to discriminate happy from all of the other emotions. We acknowledge that in a "real world" setting, emotion perception may be considerably more complex, especially when there may be multiple faces in view, all dynamically changing, and all competing with other attentional demands. Some evidence suggests that dynamic facial displays may engage additional brain processes compared to static expressions. Future work in this area might benefit from the inclusion of more complex types of facial stimuli, including dynamic expressions embedded with other competing stimuli.

Recent work has also examined similar effects in patients with sleep disorders. For instance, our findings are highly similar to a recent study that showed that patients with either insomnia or sleep apnea showed reduced accuracy for recognizing happy and sad faces, but not for other emotions such as anger, anxiety, fear, or disgust (Cronlein et al., 2016). This raises the possibility that chronic sleep problems may lead to alterations in emotional processing that are similar to that produced by experimental sleep deprivation and could have implications for many social interactions and work contexts. However, in a separate study, patients with psychophysiological insomnia showed no difference from healthy controls in their ability to accurately classify expressions of fear, anger, sadness, and happiness, although they did show lower intensity ratings for sad and fear expressions (Kyle et al., 2014). However, with a much smaller sample size in the latter study, it is possible that the failure to find differences between insomnia and controls may have been due to lack of power. Further research will be needed to clarify this issue in patient populations.

Here, we showed that while social-affiliative emotional expressions were adversely affected by sleep loss, those expressions communicating potentially hazardous conditions (i.e., anger, fear, surprise, and disgust) were sustained. These findings are consistent with a large literature suggesting that the human brain is hardwired to respond to emotional face stimuli and is particularly sensitive to facial cues communicating threat (Green and Phillips, 2004; Killgore et al., 2013; West et al., 2011). From the standpoint of short-term survival, it would make sense that crucial threat detection systems would be more robust against temporary sleep loss than those involved in social affiliation. While information about social and affiliative conditions is important for long-term survival and reproduction, it is probably less critical to the immediate survival of the individual during acute periods of potential danger. Sleep deprivation may place a person at heightened risk when functioning in a threatening environment because it degrades reaction time, cognitive processing, and physical capacities (Durmer and Dinges, 2005; Killgore, 2010). When compromised by sleep loss, survival would be most assured if an individual was able to sustain accurate or enhanced recognition of cues reflecting potential danger (e.g., a face expressing anger, fear, surprise, or disgust). In fact, during periods when physical and mental capacities have been impaired by sleep deprivation, it might actually be advantageous to err on the side of caution with regard to interpreting the meaning of emotional expressions, especially non-threat-related social emotions such as happiness and sadness. These emotional expressions essentially communicate that it is acceptable to lower one's defensive posture because it is safe to affiliate or socially appropriate to show empathy. However, it is not inconceivable that during periods of sleep deprivation, it may actually confer a survival advantage to misinterpret social/affiliative facial expressions of happiness or sadness as something a bit more ominous—just to be on the safe side. In other words, during situations when one's cognitive performance is compromised through sleep loss, the consequences of misreading a safe face as threatening are probably far less grave than misreading a threatening face as safe. Such a perspective is also consistent with other evidence suggesting that sleep deprived people are more likely to violate their own moral personal beliefs (Killgore et al., 2007), report lower levels of empathy (Killgore et al., 2008), and are more likely to feel anxious and suspicious (Kahn-Greene et al., 2007) than when not sleep deprived.

The aforementioned perspective is consistent with recent findings showing that sleep deprivation increases the general tendency to perceive facial expressions as "threatening" in appearance (Goldstein-Piekarski et al., 2015). All things being equal, a sleep-deprived individual shows a lower threshold for interpreting a face as potentially threatening than when normally rested. Moreover, this sleep-loss induced drop in the threshold for threat perception corresponds with changes in brain activation within a visero-sensory network comprising the dorsal anterior cingulate cortex and insula (Goldstein-Piekarski et al., 2015). These findings suggest that sleep deprivation lowers the threshold of interoceptive sensitivity of brain homeostatic systems, enhancing the threat-response system and affecting the interpretation of pro-social and antisocial cues (Goldstein-Piekarski et al., 2015; Simon et al., 2015). Findings from the animal literature are also supportive of this model, suggesting that deprivation of rapid eye movement sleep in pregnant rats increases defensive aggression and reduces the threshold for responding to potentially hostile stimuli (Pires et al., 2015). Overall, our findings are consistent with these data in animals and humans, suggesting that the accuracy of detecting facial displays that communicate clear and present danger may be more robust against sleep deprivation than those involved in detecting social and affiliative emotions.

Several limitations should be considered when interpreting these findings. First, we did not collect data on emotional intensity ratings in this study and it is possible that participants perceived the threat related stimuli as more arousing overall, and such increased arousal may have in turn led to greater accuracy in responding to the threat-related faces during periods of decreased alertness and vigilance. However, several studies have shown that sleep deprivation generally leads to a decrease rather than increase in the perceived intensity of emotional ratings on faces (Kyle et al., 2014; Van Der Helm et al., 2010), suggesting that increased arousal is an unlikely explanation. Second, we did not compare our data to a non-sleep deprived control group over the same time period, so it is possible that similar effects would emerge following repeated testing with the stimuli. It is therefore impossible to rule out effects of learning or carry-over effects due to repeated exposure to the stimuli across multiple sessions. The effects of learning and repeated exposure to the stimuli may have offset deficits in accurate perception of some faces, such as surprise, fear, and anger. This alternate explanation might even suggest that the threat-related emotions are more easily learned than the affiliative emotions, making them somewhat resistant to degradation by sleep deprivation. Furthermore, the fact that accuracy for fear faces was improved after the recovery night, suggests that learning/consolidation for fearful expressions may have been suppressed during sleep deprivation but re-emerged following a night of recovery sleep. In order to fully clarify the combined effects of sleep deprivation and learning, future work will need to compare repeated administrations of these stimuli to a non-sleep deprived control group. Third, the version of the task we used allowed the faces to be displayed for up to five seconds awaiting a response. Thus, there was no significant time pressure and it is possible that different results would have been obtained if the stimuli had been shown for only a fraction of a second or if a specific time pressure had been imposed. Similarly, our methods only allowed collection of accuracy data (i.e., no reaction time data were collected). However, accuracy data may not be sufficient to understand the effects of sleep deprivation on face processing, as it is well established that sleep loss significantly slows general reaction time. These questions will need to be addressed in future research. Fourth, while we utilized a set of standard basic emotional faces that have been well validated (Ekman and Friesen, 1976), it is possible that different effects would have been obtained if a more comprehensive set of emotions or expression stimuli were tested. For instance, the use of a single poser (a middle aged Caucasian male) reduces the generalizability of the findings and may not reveal the complexities of emotional face processing that would emerge with a more diverse set of stimuli in terms of age, gender, and race. At present, this study reflects the largest set of discrete emotional expressions yet tested in a laboratory sleep deprivation experiment, but more remains to be done. While our study and most others have only included a single emotion expression of happiness, emerging research suggests that there may be a number of potential positive emotions with corresponding facial displays (e.g., amusement, awe, pride, etc.) that have been relatively unexplored until recently (Keltner and Shiota, 2003; Mortillaro et al., 2011; Shiota et al., 2003). Future work should examine the effects of insufficient sleep on the perception of these other positive emotions as well. Fifth, the time of day for the testing sessions could have affected the results. Specifically, baseline and post-recovery sessions just after noon, whereas the sleep deprived session occurred early in the morning at around 0630. While this was done to maximize potential deficits in order to determine the practical effects on real-life performance, it makes it impossible to disentangle the individual effects of sleep deprivation and the circadian rhythm of performance. Future work will need to explore this effect under conditions that allow these two influences to be disambiguated. Finally, the present findings may be limited in ecological validity, as they involved fairly artificial two-dimensional displays of black and white static photographs of morphed faces. It will be important to investigate whether similar findings are also observed with more externally valid stimuli, such as dynamic displays of emotion, more diverse ethnic identities, and the incorporation of body posture and contextual cues. Despite the aforementioned limitations, the present study provides the largest sample using the broadest range of emotional facial expressions published to date to examine the effects of sleep deprivation on emotional recognition accuracy. Our findings are consistent with an adaptive survival model suggesting that during periods of sleep deprivation the brain preserves those cognitive-affective processes most relevant to threat detection at the expense of those that do not contribute to short-term survival.

## Conflicts of interest

The authors have no conflicts of interest to declare.

## Acknowledgments

This research was supported in part by an appointment to the Knowledge Preservation Program at the Walter Reed Army Institute of Research administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and USAMRMC. The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of the Army, the Department of Defense, the U.S. Government, or any of the institutions with which the authors are affiliated.

## References

Alves, N.T., Aznar-Casanova, J.A., Fukusima, S.S., 2009. Patterns of brain asymmetry in the perception of positive and negative facial expressions. Laterality 14, 256–272.
Blair, R.J., 2003. Facial expressions, their communicatory functions and neuro-cognitive substrates. Philos. Trans. R. Soc. Lond. B: Biol. Sci. 358, 561–572.
Blasi, G., Hariri, A.R., Alce, G., et al., 2009. Preferential amygdala reactivity to the negative assessment of neutral faces. Biol. Psychiatry 66, 847–853.
Cote, K.A., Mondloch, C.J., Sergeeva, V., Taylor, M., Semplonius, T., 2014. Impact of total sleep deprivation on behavioural neural processing of emotionally expressive faces. Exp. Brain Res. 232, 1429–1442.
Cronlein, T., Langguth, B., Eichhammer, P., Busch, V., 2016. Impaired recognition of facially expressed emotions in different groups of patients with sleep disorders. PLoS One 11, e0152754.
Durmer, J.S., Dinges, D.F., 2005. Neurocognitive consequences of sleep deprivation. Semin. Neurol. 25, 117–129.
Ekman, P., 1992. Are there basic emotions? Psychol. Rev. 99, 550–553.
Ekman, P., Friesen, W.V., 1976. Pictures of Facial Affect. Consulting Psychologists Press, Palo Alto, CA.
Goldstein-Piekarski, A.N., Greer, S.M., Saletin, J.M., Walker, M.P., 2015. Sleep deprivation impairs the human central and peripheral nervous system

discrimination of social threat. J. Neurosci. 35, 10135–10145.
Green, M.J., Phillips, M.L., 2004. Social threat perception and the evolution of paranoia. Neurosci. Biobehav. Rev. 28, 333–342.
Gujar, N., Yoo, S.S., Hu, P., Walker, M.P., 2011. Sleep deprivation amplifies reactivity of brain reward networks, biasing the appraisal of positive emotional experiences. J. Neurosci. 31, 4466–4474.
Huck, N.O., Mcbride, S.A., Kendall, A.P., Grugle, N.L., Killgore, W.D.S., 2008. The effects of modafinil, caffeine, and dextroamphetamine on judgments of simple versus complex emotional expressions following sleep deprivation. Int. J. Neurosci. 118, 487–502.
Kahn-Greene, E.T., Killgore, D.B., Kamimori, G.H., Balkin, T.J., Killgore, W.D.S., 2007. The effects of sleep deprivation on symptoms of psychopathology in healthy adults. Sleep Med. 8, 215–221.
Keltner, D., Shiota, M.N., 2003. New displays and new emotions: a commentary on Rozin and Cohen. Emotion 3, 86–91, (discussion 92-96).
Killgore, W.D.S., 2010. Effects of sleep deprivation on cognition. Prog. Brain Res. 185, 105–129.
Killgore, W.D.S., Kahn-Greene, E.T., Lipizzi, E.L., Newman, R.A., Kamimori, G.H., Balkin, T.J., 2008. Sleep deprivation reduces perceived emotional intelligence and constructive thinking skills. Sleep Med. 9, 517–526.
Killgore, W.D.S., Killgore, D.B., Day, L.M., Li, C., Kamimori, G.H., Balkin, T.J., 2007. The effects of 53 h of sleep deprivation on moral judgment. Sleep 30, 345–352.
Killgore, W.D.S., Schwab, Z.J., Tkachenko, O., et al., 2013. Emotional intelligence correlates with functional responses to dynamic changes in facial trustworthiness. Soc. Neurosci. 8, 334–346.
Kyle, S.D., Beattie, L., Spiegelhalder, K., Rogers, Z., Espie, C.A., 2014. Altered emotion perception in insomnia disorder. Sleep 37, 775–783.
Mortillaro, M., Mehu, M., Scherer, K., 2011. Subtly different positive emotions can be distinguished by their facial expressions. Social. Psychol. Pers. Sci. 2, 262–271.
Pires, G.N., Tufik, S., Andersen, M.L., 2015. Effects of REM sleep restriction during pregnancy on rodent maternal behavior. Rev. Bras. Psiquiatr. 37, 303–309.
Shiota, M.N., Campos, B., Keltner, D., 2003. The faces of positive emotion: prototype displays of awe, amusement, and pride. Ann. N. Y. Acad. Sci. 1000, 296–299.
Simon, E.B., Oren, N., Sharon, H., et al., 2015. Losing neutrality: the neural basis of impaired emotional control without sleep. J. Neurosci. 35, 13194–13205.
Van Der Helm, E., Gujar, N., Walker, M.P., 2010. Sleep deprivation impairs the accurate recognition of human emotions. Sleep 33, 335–342.
Walker, M.P., 2009. The role of sleep in cognition and emotion. Ann. N. Y. Acad. Sci. 1156, 168–197.
Walker, M.P., Van Der Helm, E., 2009. Overnight therapy? The role of sleep in emotional brain processing. Psychol. Bull. 135, 731–748.
Wells, L.J., Gillespie, S.M., Rotshtein, P., 2016. Identification of emotional facial expressions: effects of expression, intensity, and sex on eye gaze. PLoS One 11, e0168307.
West, G.L., Anderson, A.A., Ferber, S., Pratt, J., 2011. Electrophysiological evidence for biased competition in V1 for fear expressions. J. Cogn. Neurosci. 23, 3410–3418.
Yoo, S.-S., Gujar, N., Hu, P., Jolesz, F.A., Walker, M.P., 2007. The human emotional brain without sleep - a prefrontal amygdala disconnect. Curr. Biol.: CB 17, R877–R878.

# Active learning increases student performance in science, engineering, and mathematics

Scott Freeman[a,1], Sarah L. Eddy[a], Miles McDonough[a], Michelle K. Smith[b], Nnadozie Okoroafor[a], Hannah Jordt[a], and Mary Pat Wenderoth[a]

[a]Department of Biology, University of Washington, Seattle, WA 98195; and [b]School of Biology and Ecology, University of Maine, Orono, ME 04469

Edited* by Bruce Alberts, University of California, San Francisco, CA, and approved April 15, 2014 (received for review October 8, 2013)

**To test the hypothesis that lecturing maximizes learning and course performance, we metaanalyzed 225 studies that reported data on examination scores or failure rates when comparing student performance in undergraduate science, technology, engineering, and mathematics (STEM) courses under traditional lecturing versus active learning. The effect sizes indicate that on average, student performance on examinations and concept inventories increased by 0.47 SDs under active learning ($n$ = 158 studies), and that the odds ratio for failing was 1.95 under traditional lecturing ($n$ = 67 studies). These results indicate that average examination scores improved by about 6% in active learning sections, and that students in classes with traditional lecturing were 1.5 times more likely to fail than were students in classes with active learning. Heterogeneity analyses indicated that both results hold across the STEM disciplines, that active learning increases scores on concept inventories more than on course examinations, and that active learning appears effective across all class sizes—although the greatest effects are in small ($n \leq 50$) classes. Trim and fill analyses and fail-safe $n$ calculations suggest that the results are not due to publication bias. The results also appear robust to variation in the methodological rigor of the included studies, based on the quality of controls over student quality and instructor identity. This is the largest and most comprehensive metaanalysis of undergraduate STEM education published to date. The results raise questions about the continued use of traditional lecturing as a control in research studies, and support active learning as the preferred, empirically validated teaching practice in regular classrooms.**

constructivism | undergraduate education | evidence-based teaching | scientific teaching

Lecturing has been the predominant mode of instruction since universities were founded in Western Europe over 900 y ago (1). Although theories of learning that emphasize the need for students to construct their own understanding have challenged the theoretical underpinnings of the traditional, instructor-focused, "teaching by telling" approach (2, 3), to date there has been no quantitative analysis of how constructivist versus exposition-centered methods impact student performance in undergraduate courses across the science, technology, engineering, and mathematics (STEM) disciplines. In the STEM classroom, should we ask or should we tell?

Addressing this question is essential if scientists are committed to teaching based on evidence rather than tradition (4). The answer could also be part of a solution to the "pipeline problem" that some countries are experiencing in STEM education: For example, the observation that less than 40% of US students who enter university with an interest in STEM, and just 20% of STEM-interested underrepresented minority students, finish with a STEM degree (5).

To test the efficacy of constructivist versus exposition-centered course designs, we focused on the design of class sessions—as opposed to laboratories, homework assignments, or other exercises. More specifically, we compared the results of experiments that documented student performance in courses with at least some active learning versus traditional lecturing, by metaanalyzing 225 studies in the published and unpublished literature. The active learning interventions varied widely in intensity and implementation, and included approaches as diverse as occasional group problem-solving, worksheets or tutorials completed during class, use of personal response systems with or without peer instruction, and studio or workshop course designs. We followed guidelines for best practice in quantitative reviews (*SI Materials and Methods*), and evaluated student performance using two outcome variables: (*i*) scores on identical or formally equivalent examinations, concept inventories, or other assessments; or (*ii*) failure rates, usually measured as the percentage of students receiving a D or F grade or withdrawing from the course in question (DFW rate).

The analysis, then, focused on two related questions. Does active learning boost examination scores? Does it lower failure rates?

## Results

The overall mean effect size for performance on identical or equivalent examinations, concept inventories, and other assessments was a weighted standardized mean difference of 0.47 ($Z$ = 9.781, $P << 0.001$)—meaning that on average, student performance increased by just under half a SD with active learning compared with lecturing. The overall mean effect size for failure rate was an odds ratio of 1.95 ($Z$ = 10.4, $P << 0.001$). This odds ratio is equivalent to a risk ratio of 1.5, meaning that on average, students in traditional lecture courses are 1.5 times more likely to fail than students in courses with active learning. Average failure rates were 21.8% under active learning but 33.8% under traditional lecturing—a difference that represents a 55% increase (Fig. 1 and Fig. S1).

### Significance

**The President's Council of Advisors on Science and Technology has called for a 33% increase in the number of science, technology, engineering, and mathematics (STEM) bachelor's degrees completed per year and recommended adoption of empirically validated teaching practices as critical to achieving that goal. The studies analyzed here document that active learning leads to increases in examination performance that would raise average grades by a half a letter, and that failure rates under traditional lecturing increase by 55% over the rates observed under active learning. The analysis supports theory claiming that calls to increase the number of students receiving STEM degrees could be answered, at least in part, by abandoning traditional lecturing in favor of active learning.**

Author contributions: S.F. and M.P.W. designed research; S.F., M.M., M.K.S., N.O., H.J., and M.P.W. performed research; S.F. and S.L.E. analyzed data; and S.F., S.L.E., M.M., M.K.S., N.O., H.J., and M.P.W. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

See Commentary on page 8319.

[1]To whom correspondence should be addressed. E-mail: srf991@u.washington.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1319030111/-/DCSupplemental.

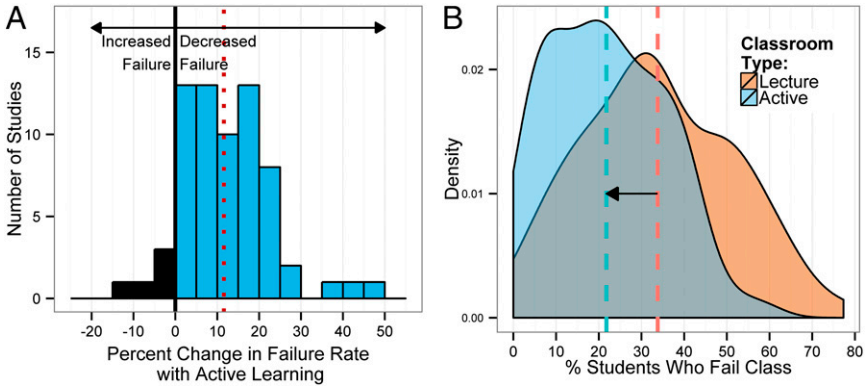


**Fig. 1.** Changes in failure rate. (*A*) Data plotted as percent change in failure rate in the same course, under active learning versus lecturing. The mean change (12%) is indicated by the dashed vertical line. (*B*) Kernel density plots of failure rates under active learning and under lecturing. The mean failure rates under each classroom type (21.8% and 33.8%) are shown by dashed vertical lines.

Heterogeneity analyses indicated no statistically significant variation among experiments based on the STEM discipline of the course in question, with respect to either examination scores (Fig. 2*A*; $Q = 910.537$, df = 7, $P = 0.160$) or failure rates (Fig. 2*B*; $Q = 11.73$, df = 6, $P = 0.068$). In every discipline with more than 10 experiments that met the admission criteria for the meta-analysis, average effect sizes were statistically significant for either examination scores or failure rates or both (Fig. 2, Figs. S2 and S3, and Tables S1*A* and S2*A*). Thus, the data indicate that active learning increases student performance across the STEM disciplines.

For the data on examinations and other assessments, a heterogeneity analysis indicated that average effect sizes were lower when the outcome variable was an instructor-written course examination as opposed to performance on a concept inventory (Fig. 3*A* and Table S1*B*; $Q = 10.731$, df = 1, $P << 0.001$). Although student achievement was higher under active learning for both types of assessments, we hypothesize that the difference in gains for examinations versus concept inventories may be due to the two types of assessments testing qualitatively different cognitive skills. This explanation is consistent with previous research indicating that active learning has a greater impact on student mastery of higher- versus lower-level cognitive skills (6–9), and the recognition that most concept inventories are designed to diagnose known misconceptions, in contrast to course examinations that emphasize content mastery or the ability to solve quantitative problems (10). Most concept inventories also undergo testing for validity, reliability, and readability.

Heterogeneity analyses indicated significant variation in terms of course size, with active learning having the highest impact on courses with 50 or fewer students (Fig. 3*B* and Table S1*C*; $Q = 6.726$, df = 2, $P = 0.035$; Fig. S4). Effect sizes were statistically significant for all three categories of class size, however, indicating that active learning benefitted students in medium (51–110 students) or large (>110 students) class sizes as well.

When we metaanalyzed the data by course type and course level, we found no statistically significant difference in active learning's effect size when comparing (*i*) courses for majors versus nonmajors ($Q = 0.045$, df = 1, $P = 0.883$; Table S1*D*), or (*ii*) introductory versus upper-division courses ($Q = 0.046$, df = 1, $P = 0.829$; Tables S1*E* and S2*D*).

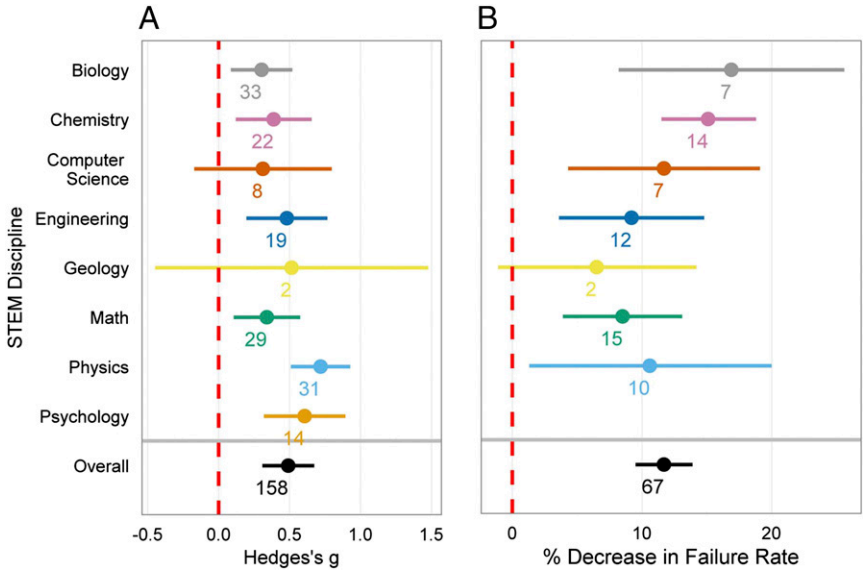


**Fig. 2.** Effect sizes by discipline. (*A*) Data on examination scores, concept inventories, or other assessments. (*B*) Data on failure rates. Numbers below data points indicate the number of independent studies; horizontal lines are 95% confidence intervals.

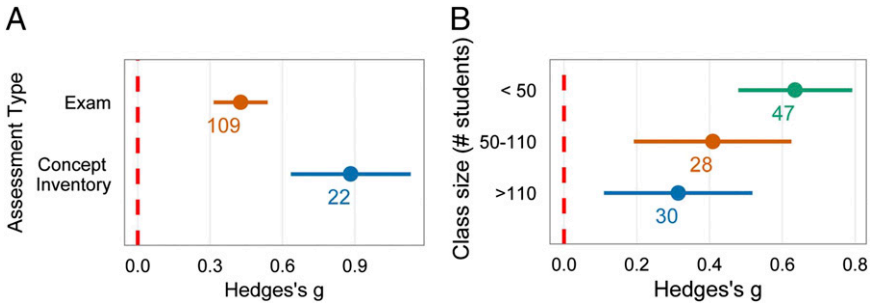


**Fig. 3.** Heterogeneity analyses for data on examination scores, concept inventories, or other assessments. (*A*) By assessment type—concept inventories versus examinations. (*B*) By class size. Numbers below data points indicate the number of independent studies; horizontal lines are 95% confidence intervals.

To evaluate how confident practitioners can be about these conclusions, we performed two types of analyses to assess whether the results were compromised by publication bias, i.e., the tendency for studies with low effect sizes to remain unpublished. We calculated fail-safe numbers indicating how many missing studies with an effect size of 0 would have to be published to reduce the overall effect sizes of 0.47 for examination performance and 1.95 for failure rate to preset levels that would be considered small or moderate—in this case, 0.20 and 1.1, respectively. The fail-safe numbers were high: 114 studies on examination performance and 438 studies on failure rate (*SI Materials and Methods*). Analyses of funnel plots (Fig. S5) also support a lack of publication bias (*SI Materials and Methods*).

To assess criticisms that the literature on undergraduate STEM education is difficult to interpret because of methodological shortcomings (e.g., ref. 11), we looked for heterogeneity in effect sizes for the examination score data, based on whether experiments did or did not meet our most stringent criteria for student and instructor equivalence. We created four categories to characterize the quality of the controls over student equivalence in the active learning versus lecture treatments (*SI Materials and Methods*), and found that there was no heterogeneity based on methodological quality ($Q = 2.097$, df = 3, $P = 0.553$): Experiments where students were assigned to treatments at random produced results that were indistinguishable from three types of quasirandomized designs (Table 1). Analyzing variation with respect to controls over instructor identity also produced no evidence of heterogeneity ($Q = 0.007$, df = 1, $P = 0.934$): More poorly controlled studies, with different instructors in the two treatment groups or with no data provided on instructor equivalence, gave equivalent results to studies with identical or randomized instructors in the two treatments (Table 1). Thus, the overall effect size for examination data appears robust to variation in the methodological rigor of published studies.

## Discussion

The data reported here indicate that active learning increases examination performance by just under half a SD and that lecturing increases failure rates by 55%. The heterogeneity analyses indicate that (*i*) these increases in achievement hold across all of the STEM disciplines and occur in all class sizes, course types, and course levels; and (*ii*) active learning is particularly beneficial in small classes and at increasing performance on concept inventories.

Although this is the largest and most comprehensive meta-analysis of the undergraduate STEM education literature to date, the weighted, grand mean effect size of 0.47 reported here is almost identical to the weighted, grand-mean effect sizes of 0.50 and 0.51 published in earlier metaanalyses of how alternatives to traditional lecturing impact undergraduate course performance in subsets of STEM disciplines (11, 12). Thus, our results are consistent with previous work by other investigators.

The grand mean effect sizes reported here are subject to important qualifications, however. For example, because struggling students are more likely to drop courses than high-achieving students, the reductions in withdrawal rates under active learning that are documented here should depress average scores on assessments—meaning that the effect size of 0.47 for examination and concept inventory scores may underestimate active learning's actual impact in the studies performed to date (*SI Materials and Methods*). In contrast, it is not clear whether effect sizes of this magnitude would be observed if active learning approaches were to become universal. The instructors who implemented active learning in these studies did so as volunteers. It is an open question whether student performance would increase as much if all faculty were required to implement active learning approaches.

Assuming that other instructors implement active learning and achieve the average effect size documented here, what would

**Table 1. Comparing effect sizes estimated from well-controlled versus less-well-controlled studies**

| | | | | 95% confidence interval | |
|---|---|---|---|---|---|
| Type of control | *n* | Hedges's *g* | SE | Lower limit | Upper limit |
| For student equivalence | | | | | |
| Quasirandom—no data on student equivalence | 39 | 0.467 | 0.102 | 0.268 | 0.666 |
| Quasirandom—no statistical difference in prescores on assessment used for effect size | 51 | 0.534 | 0.089 | 0.359 | 0.709 |
| Quasirandom—no statistical difference on metrics of academic ability/preparedness | 51 | 0.362 | 0.092 | 0.181 | 0.542 |
| Randomized assignment or crossover design | 16 | 0.514 | 0.098 | 0.322 | 0.706 |
| For instructor equivalence | | | | | |
| No data, or different instructors | 59 | 0.472 | 0.081 | 0.313 | 0.631 |
| Identical instructor, randomized assignment, or ≥3 instructors in each treatment | 99 | 0.492 | 0.071 | 0.347 | 0.580 |

a shift of 0.47 SDs in examination and concept inventory scores mean to their students?

*i*) Students performing in the 50th percentile of a class based on traditional lecturing would, under active learning, move to the 68th percentile of that class (13)—meaning that instead of scoring better than 50% of the students in the class, the same individual taught with active learning would score better than 68% of the students being lectured to.

*ii*) According to an analysis of examination scores in three introductory STEM courses (*SI Materials and Methods*), a change of 0.47 SDs would produce an increase of about 6% in average examination scores and would translate to a 0.3 point increase in average final grade. On a letter-based system, medians in the courses analyzed would rise from a B− to a B or from a B to a B+.

The result for undergraduate STEM courses can also be compared with the impact of educational interventions at the precollege level. A recent review of educational interventions in the K–12 literature reports a mean effect size of 0.39 when impacts are measured with researcher-developed tests, analogous to the examination scores analyzed here, and a mean effect size of 0.24 for narrow-scope standardized tests, analogous to the concept inventories analyzed here (14). Thus, the effect size of active learning at the undergraduate level appears greater than the effect sizes of educational innovations in the K–12 setting, where effect sizes of 0.20 or even smaller may be considered of policy interest (14).

There are also at least two ways to view an odds ratio of 1.95 for the risk of failing a STEM course:

*i*) If the experiments analyzed here had been conducted as randomized controlled trials of medical interventions, they may have been stopped for benefit—meaning that enrolling patients in the control condition might be discontinued because the treatment being tested was clearly more beneficial. For example, a recent analysis of 143 randomized controlled medical trials that were stopped for benefit found that they had a median relative risk of 0.52, with a range of 0.22 to 0.66 (15). In addition, best-practice directives suggest that data management committees may allow such studies to stop for benefit if interim analyses have large sample sizes and $P$ values under 0.001 (16). Both criteria were met for failure rates in the education studies we analyzed: The average relative risk was 0.64 and the $P$ value on the overall odds ratio was $<< 0.001$. Any analogy with biomedical trials is qualified, however, by the lack of randomized designs in studies that included data on failure rates.

*ii*) There were 29,300 students in the 67 lecturing treatments with data on failure rates. Given that the raw failure rate in this sample averaged 33.8% under traditional lecturing and 21.8% under active learning, the data suggest that 3,516 fewer students would have failed these STEM courses under active learning. Based on conservative assumptions (*SI Materials and Methods*), this translates into over US$3,500,000 in saved tuition dollars for the study population, had all students been exposed to active learning. If active learning were implemented widely, the total tuition dollars saved would be orders of magnitude larger, given that there were 21 million students enrolled in US colleges and universities alone in 2010, and that about a third of these students intended to major in STEM fields as entering freshmen (17, 18).

Finally, increased grades and fewer failures should make a significant impact on the pipeline problem. For example, the 2012 President's Council of Advisors on Science and Technology report calls for an additional one million STEM majors in the United States in the next decade—requiring a 33% increase from the current annual total—and notes that simply increasing the current STEM retention rate of 40% to 50% would meet three-quarters of that goal (5). According to a recent cohort study from the National Center for Education Statistics (19), there are gaps of 0.5 and 0.4 in the STEM-course grade point averages (GPAs) of first-year bachelor's and associate's degree students, respectively, who end up leaving versus persisting in STEM programs. A 0.3 "bump" in average grades with active learning would get the "leavers" close to the current performance level of "persisters." Other analyses of students who leave STEM majors indicate that increased passing rates, higher grades, and increased engagement in courses all play a positive role in retention (20–22).

In addition to providing evidence that active learning can improve undergraduate STEM education, the results reported here have important implications for future research. The studies we metaanalyzed represent the first-generation of work on undergraduate STEM education, where researchers contrasted a diverse array of active learning approaches and intensities with traditional lecturing. Given our results, it is reasonable to raise concerns about the continued use of traditional lecturing as a control in future experiments. Instead, it may be more productive to focus on what we call "second-generation research": using advances in educational psychology and cognitive science to inspire changes in course design (23, 24), then testing hypotheses about which type of active learning is most appropriate and efficient for certain topics or student populations (25). Second-generation research could also explore which aspects of instructor behavior are most important for achieving the greatest gains with active learning, and elaborate on recent work indicating that underprepared and underrepresented students may benefit most from active methods. In addition, it will be important to address questions about the intensity of active learning: Is more always better? Although the time devoted to active learning was highly variable in the studies analyzed here, ranging from just 10–15% of class time being devoted to clicker questions to lecture-free "studio" environments, we were not able to evaluate the relationship between the intensity (or type) of active learning and student performance, due to lack of data (*SI Materials and Methods*).

As research continues, we predict that course designs inspired by second-generation studies will result in additional gains in student achievement, especially when the types of active learning interventions analyzed here—which focused solely on in-class innovations—are combined with required exercises that are completed outside of formal class sessions (26).

Finally, the data suggest that STEM instructors may begin to question the continued use of traditional lecturing in everyday practice, especially in light of recent work indicating that active learning confers disproportionate benefits for STEM students from disadvantaged backgrounds and for female students in male-dominated fields (27, 28). Although traditional lecturing has dominated undergraduate instruction for most of a millennium and continues to have strong advocates (29), current evidence suggests that a constructivist "ask, don't tell" approach may lead to strong increases in student performance—amplifying recent calls from policy makers and researchers to support faculty who are transforming their undergraduate STEM courses (5, 30).

## Materials and Methods

To create a working definition of active learning, we collected written definitions from 338 audience members, before biology departmental seminars on active learning, at universities throughout the United States and Canada. We then coded elements in the responses to create the following consensus definition:

> Active learning engages students in the process of learning through activities and/or discussion in class, as opposed to passively listening

to an expert. It emphasizes higher-order thinking and often involves group work. (See also ref. 31, p. iii).

Following Bligh (32), we defined traditional lecturing as "...continuous exposition by the teacher." Under this definition, student activity was assumed to be limited to taking notes and/or asking occasional and unprompted questions of the instructor.

**Literature Search.** We searched the gray literature, primarily in the form of unpublished dissertations and conference proceedings, in addition to peer-reviewed sources (33, 34) for studies that compared student performance in undergraduate STEM courses under traditional lecturing versus active learning. We used four approaches (35) to find papers for consideration: hand-searching every issue in 55 STEM education journals from June 1, 1998 to January 1, 2010 (Table S3), searching seven online databases using an array of terms, mining reviews and bibliographies (*SI Materials and Methods*), and "snowballing" from references in papers admitted to the study (*SI Materials and Methods*). We had no starting time limit for admission to the study; the ending cutoff for consideration was completion or publication before January 1, 2010.

**Criteria for Admission.** As recommended (36), the criteria for admission to the coding and final data analysis phases of the study were established at the onset of the work and were not altered. We coded studies that (*i*) contrasted traditional lecturing with any active learning intervention, with total class time devoted to each approach not differing by more than 30 min/wk; (*ii*) occurred in the context of a regularly scheduled course for undergraduates; (*iii*) were largely or solely limited to changes in the conduct of the regularly scheduled class or recitation sessions; (*iv*) involved a course in astronomy, biology, chemistry, computer science, engineering, geology, mathematics, natural resources or environmental science, nutrition or food science, physics, psychology, or statistics; and (*v*) included data on some aspect of student academic performance.

Note that criterion *i* yielded papers representing a wide array of active learning activities, including vaguely defined "cooperative group activities in class," in-class worksheets, clickers, problem-based learning (PBL), and studio classrooms, with intensities ranging from 10% to 100% of class time (*SI Materials and Methods*). Thus, this study's intent was to evaluate the average effect of any active learning type and intensity contrasted with traditional lecturing.

The literature search yielded 642 papers that appeared to meet these five criteria and were subsequently coded by at least one of the authors.

**Coding.** All 642 papers were coded by one of the authors (S.F.) and 398 were coded independently by at least one other member of the author team (M.M., M.S., M.P.W., N.O., or H.J.). The 244 "easy rejects" were excluded from the study after the initial coder (S.F.) determined that they clearly did not meet one or more of the five criteria for admission; a post hoc analysis suggested that the easy rejects were justified (*SI Materials and Methods*).

The two coders met to review each of the remaining 398 papers and reach consensus (37, 38) on

*i*) The five criteria listed above for admission to the study;
*ii*) Examination equivalence—meaning that the assessment given to students in the lecturing and active learning treatment groups had to be identical, equivalent as judged by at least one third-party observer recruited by the authors of the study in question but blind to the hypothesis being tested, or comprising questions drawn at random from a common test bank;
*iii*) Student equivalence—specifically whether the experiment was based on randomization or quasirandomization among treatments and, if quasirandom, whether students in the lecture and active learning treatments were statistically indistinguishable in terms of (*a*) prior general academic performance (usually measured by college GPA at the time of entering the course, Scholastic Aptitude Test, or American College Testing scores), or (*b*) pretests directly relevant to the topic in question;
*iv*) Instructor equivalence—meaning whether the instructors in the lecture and active learning treatments were identical, randomly assigned, or consisted of a group of three or more in each treatment; and
*v*) Data that could be used for computing an effect size.

To reduce or eliminate pseudoreplication, the coders also annotated the effect size data using preestablished criteria to identify and report effect sizes only from studies that represented independent courses and populations reported. If the data reported were from iterations of the same course at the same institution, we combined data recorded for more than one control and/or treatment group from the same experiment. We also combined data from multiple outcomes from the same study (e.g., a series of equivalent midterm examinations) (*SI Materials and Methods*). Coders also extracted data on class size, course type, course level, and type of active learning, when available.

Criteria *iii* and *iv* were meant to assess methodological quality in the final datasets, which comprised 158 independent comparisons with data on student examination performance and 67 independent comparisons with data on failure rates. The data analyzed and references to the corresponding papers are archived in Table S4.

**Data Analysis.** Before analyzing the data, we inspected the distribution of class sizes in the study and binned this variable as small, medium, and large (*SI Materials and Methods*). We also used established protocols (38, 39) to combine data from multiple treatments/controls and/or data from multiple outcomes, and thus produce a single pairwise comparison from each independent course and student population in the study (*SI Materials and Methods*).

The data we analyzed came from two types of studies: (*i*) randomized trials, where each student was randomly placed in a treatment; and (*ii*) quasirandom designs where students self-sorted into classes, blind to the treatment at the time of registering for the class. It is important to note that in the quasirandom experiments, students were assigned to treatment as a group, meaning that they are not statistically independent samples. This leads to statistical problems: The number of independent data points in each treatment is not equal to the number of students (40). The element of nonindependence in quasirandom designs can cause variance calculations to underestimate the actual variance, leading to overestimates for significance levels and for the weight that each study is assigned (41). To correct for this element of nonindependence in quasirandom studies, we used a cluster adjustment calculator in Microsoft Excel based on methods developed by Hedges (40) and implemented in several recent metaanalyses (42, 43). Adjusting for clustering in our data required an estimate of the intraclass correlation coefficient (ICC). None of our studies reported ICCs, however, and to our knowledge, no studies have reported an ICC in college-level STEM courses. Thus, to obtain an estimate for the ICC, we turned to the K–12 literature. A recent paper reviewed ICCs for academic achievement in mathematics and reading for a national sample of K–12 students (44). We used the mean ICC reported for mathematics (0.22) as a conservative estimate of the ICC in college-level STEM classrooms. Note that although the cluster correction has a large influence on the variance for each study, it does not influence the effect size point estimate substantially.

We computed effect sizes and conducted the metaanalysis in the Comprehensive Meta-Analysis software package (45). All reported *P* values are two-tailed, unless noted.

We used a random effects model (46, 47) to compare effect sizes. The random effect size model was appropriate because conditions that could affect learning gains varied among studies in the analysis, including the (*i*) type (e.g., PBL versus clickers), intensity (percentage of class time devoted to constructivist activities), and implementation (e.g., graded or ungraded) of active learning; (*ii*) student population; (*iii*) course level and discipline; and (*iv*) type, cognitive level, and timing—relative to the active learning exercise—of examinations or other assessments.

We calculated effect sizes as (*i*) the weighted standardized mean difference as Hedges' *g* (48) for data on examination scores, and (*ii*) the log-odds for data on failure rates. For ease of interpretation, we then converted log-odds values to odds ratio, risk ratio, or relative risk (49).

To evaluate the influence of publication bias on the results, we assessed funnel plots visually (50) and statistically (51), applied Duval and Tweedie's trim and fill method (51), and calculated fail-safe *N*s (45).

**Additional Results.** We did not insist that assessments be identical or formally equivalent if studies reported only data on failure rates. To evaluate the hypothesis that differences in failure rates recorded under traditional lecturing and active learning were due to changes in the difficulty of examinations and other course assessments, we evaluated 11 studies where failure rate data were based on comparisons in which most or all examination questions were identical. The average odds ratio for these 11 studies was 1.97 ± 0.36 (SE)—almost exactly the effect size calculated from the entire dataset.

Although we did not metaanalyze the data using "vote-counting" approaches, it is informative to note that of the studies reporting statistical tests of examination score data, 94 reported significant gains under active learning whereas only 41 did not (Table S4*A*).

Additional results from the analyses on publication bias are reported in Supporting Information.

**ACKNOWLEDGMENTS.** We thank Roddy Theobald for advice on interpreting odds ratios; the many authors who provided missing data upon request (*SI Materials and Methods*); Colleen Craig, Daryl Pedigo, and Deborah Wiegand for supplying information on examination score standard deviations and grading thresholds; Kelly Puzio and an anonymous reviewer for advice on analyzing data from quasirandom studies; and Steven Kroiss, Carl Wieman, and William Wood for comments that improved the manuscript. M.S. was supported in part by National Science Foundation Grant 0962805.

1. Brockliss L (1996) Curricula. *A History of the University in Europe*, ed de Ridder-Symoens H (Cambridge Univ Press, Cambridge, UK), Vol II, pp 565–620.
2. Piaget J (1926) *The Language and Thought of the Child* (Harcourt Brace, New York).
3. Vygotsky LS (1978) *Mind in Society* (Harvard Univ Press, Cambridge, MA).
4. Handelsman J, et al. (2004) Education. Scientific teaching. *Science* 304(5670):521–522.
5. PCAST STEM Undergraduate Working Group (2012) *Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics*, eds Gates SJ, Jr, Handelsman J, Lepage GP, Mirkin C (Office of the President, Washington).
6. Haukoos GD, Penick JE (1983) The influence of classroom climate on science process and content achievement of community college students. *J Res Sci Teach* 20(7): 629–637.
7. Martin T, Rivale SD, Diller KR (2007) Comparison of student learning in challenge-based and traditional instruction in biomedical engineering. *Ann Biomed Eng* 35(8): 1312–1323.
8. Cordray DS, Harris TR, Klein S (2009) A research synthesis of the effectiveness, replicability, and generality of the VaNTH challenge-based instructional modules in bioengineering. *J. Eng Ed* 98(4).
9. Jensen JL, Lawson A (2011) Effects of collaborative group composition and inquiry instruction on reasoning gains and achievement in undergraduate biology. *CBE Life Sci Educ* 10(1):64–73.
10. Momsen JL, Long TM, Wyse SA, Ebert-May D (2010) Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills. *CBE Life Sci Educ* 9(4): 435–440.
11. Ruiz-Primo MA, Briggs D, Iverson H, Talbot R, Shepard LA (2011) Impact of undergraduate science course innovations on learning. *Science* 331(6022):1269–1270.
12. Springer L, Stanne ME, Donovan SS (1999) Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology. *Rev Educ Res* 69(1):21–51.
13. Bowen CW (2000) A quantitative literature review of cooperative learning effects on high school and college chemistry achievement. *J Chem Educ* 77(1):116–119.
14. Lipsey MW, et al. (2012) *Translating the Statistical Representation of the Effects of Educational Interventions into Readily Interpretable Forms* (US Department of Education, Washington).
15. Montori VM, et al. (2005) Randomized trials stopped early for benefit: A systematic review. *JAMA* 294(17):2203–2209.
16. Pocock SJ (2006) Current controversies in data monitoring for clinical trials. *Clin Trials* 3(6):513–521.
17. National Center for Education Statistics (2012) *Digest of Education Statistics* (US Department of Education, Washington).
18. National Science Board (2010) *Science and Engineering Indicators 2010* (National Science Foundation, Arlington, VA).
19. National Center for Education Statistics (2012) *STEM in Postsecondary Education* (US Department of Education, Washington).
20. Seymour E, Hewitt NM (1997) *Talking About Leaving: Why Undergraduates Leave the Sciences* (Westview Press, Boulder, CO).
21. Goodman IF, et al. (2002) *Final Report of the Women's Experiences in College Engineering (WECE) Project* (Goodman Research Group, Cambridge, MA).
22. Watkins J, Mazur E (2013) Retaining students in science, technology, engineering, and mathematics (STEM) majors. *J Coll Sci Teach* 42(5):36–41.
23. Slavich GM, Zimbardo PG (2012) Transformational teaching: Theoretical underpinnings, basic principles, and core methods. *Educ Psychol Rev* 24(4):569–608.
24. Dunlosky J, Rawson KA, Marsh EJ, Nathan MJ, Willingham DT (2013) Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psych Sci Publ Int* 14(1):4–58.
25. Eddy S, Crowe AJ, Wenderoth MP, Freeman S (2013) How should we teach tree-thinking? An experimental test of two hypotheses. *Evol Ed Outreach* 6:1–11.
26. Freeman S, Haak D, Wenderoth MP (2011) Increased course structure improves performance in introductory biology. *CBE Life Sci Educ* 10(2):175–186.
27. Lorenzo M, Crouch CH, Mazur E (2006) Reducing the gender gap in the physics classroom. *Am J Phys* 74(2):118–122.
28. Haak DC, HilleRisLambers J, Pitre E, Freeman S (2011) Increased structure and active learning reduce the achievement gap in introductory biology. *Science* 332(6034): 1213–1216.
29. Burgan M (2006) In defense of lecturing. *Change* 6:31–34.
30. Henderson C, Beach A, Finkelstein N (2011) Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature. *J Res Sci Teach* 48(8): 952–984.
31. Bonwell CC, Eison JA (1991) *Active Learning: Creating Excitement in the Classroom* (George Washington Univ, Washington, DC).
32. Bligh DA (2000) *What's the Use of Lectures?* (Jossey-Bass, San Francisco).
33. Reed JG, Baxter PM (2009) Using reference databases. *The Handbook of Research Synthesis and Meta-Analysis*, eds Cooper H, Hedges LV, Valentine JC (Russell Sage Foundation, New York), pp 73–101.
34. Rothstein H, Hopewell S (2009) Grey literature. *The Handbook of Research Synthesis and Meta-Analysis*, eds Cooper H, Hedges LV, Valentine JC (Russell Sage Foundation, New York), pp 103–125.
35. White HD (2009) Scientific communication and literature retrieval. *The Handbook of Research Synthesis and Meta-Analysis*, eds Cooper H, Hedges LV, Valentine JC (Russell Sage Foundation, New York), pp 51–71.
36. Lipsey MW, Wilson DB (2001) *Practical Meta-Analysis* (Sage Publications, Thousand Oaks, CA).
37. Orwin RG, Vevea JL (2009) Evaluating coding decisions. *The Handbook of Research Synthesis and Meta-Analysis*, eds Cooper H, Hedges LV, Valentine JC (Russell Sage Foundation, New York), pp 177–203.
38. Higgins JPT, Green S, eds (2011) Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0 (The Cochrane Collaboration, Oxford). Available at www.cochrane-handbook.org. Accessed December 14, 2012.
39. Borenstein M (2009) Effect sizes for continuous data. *The Handbook of Systematic Review and Meta-Analysis*, eds Cooper H, Hedges LV, Valentine JC (Russell Sage Foundation, New York), pp 221–235.
40. Hedges LV (2007) Correcting a significance test for clustering. *J Educ Behav Stat* 32(2): 151–179.
41. Donner A, Klar N (2002) Issues in the meta-analysis of cluster randomized trials. *Stat Med* 21(19):2971–2980.
42. Davis D (2012) Multiple Comprehension Strategies Instruction (MCSI) for Improving Reading Comprehension and Strategy Outcomes in the Middle Grades. (The Campbell Collaboration, Oxford). Available at http://campbellcollaboration.org/lib/project/167/. Accessed December 10, 2013.
43. Puzio K, Colby GT (2013) Cooperative learning and literacy: A meta-analytic review. *J Res Ed Effect* 6(4):339–360.
44. Hedges LV, Hedberg EC (2007) Intraclass correlation values for planning group-randomized trials in education. *Educ Eval Policy Anal* 29:60–87.
45. Borenstein M, et al. (2006) *Comprehensive Meta-Analysis* (Biostat, Inc., Englewood, NJ).
46. Hedges LV (2009) Statistical considerations. *The Handbook of Research Synthesis and Meta-Analysis*, eds Cooper H, Hedges LV, Valentine JC (Russell Sage Foundation, New York), pp 38–47.
47. Raudenbush SW (2009) Analyzing effect sizes: Random-effects models. *The Handbook of Research Synthesis and Meta-Analysis*, eds Cooper H, Hedges LV, Valentine JC (Russell Sage Foundation, New York), pp 295–315.
48. Gurevitch J, Hedges LV (1999) Statistical issues in ecological meta-analyses. *Ecology* 80(4):1142–1149.
49. Fleiss J, Berlin JA (2009) Effect sizes for dichotomous data. *The Handbook of Research Synthesis and Meta-Analysis*, eds Cooper H, Hedges LV, Valentine JC (Russell Sage Foundation, New York), pp 237–253.
50. Greenhouse JB, Iyengar S (2009) Sensitivity analysis and diagnostics. *The Handbook of Research Synthesis and Meta-Analysis*, eds Cooper H, Hedges LV, Valentine JC (Russell Sage Foundation, New York), pp 417–433.
51. Sutton AJ (2009) Publication bias. *The Handbook of Research Synthesis and Meta-Analysis*, eds Cooper H, Hedges LV, Valentine JC (Russell Sage Foundation, New York), pp 435–452.

Journal of Experimental Psychology: Applied
2009, Vol. 15, No. 3, 243–257


1076-898X/09/$12.00 DOI: 10.1037/a0016496

# The Pretesting Effect: Do Unsuccessful Retrieval Attempts Enhance Learning?

Lindsey E. Richland
University of California, Irvine

Nate Kornell
Williams College

Liche Sean Kao
University of California, Irvine

Testing previously studied information enhances long-term memory, particularly when the information is successfully retrieved from memory. The authors examined the effect of unsuccessful retrieval attempts on learning. Participants in 5 experiments read an essay about vision. In the test condition, they were asked about embedded concepts before reading the passage; in the extended study condition, they were given a longer time to read the passage. To distinguish the effects of testing from attention direction, the authors emphasized the tested concepts in both conditions, using italics or bolded keywords or, in Experiment 5, by presenting the questions but not asking participants to answer them before reading the passage. Posttest performance was better in the test condition than in the extended study condition in all experiments—a pretesting effect—even though only items that were not successfully retrieved on the pretest were analyzed. The testing effect appears to be attributable, in part, to the role unsuccessful tests play in enhancing future learning.

*Keywords:* testing, learning, memory, retrieval

Testing has become a central issue in the current U.S. political debate concerning education. To ensure equal access to a high-quality education, operationalized as proficiency on state academic assessments (No Child Left Behind Act, 2001), educational reforms have replaced instruction—sometimes several weeks' worth each year—with standardized testing in an effort to monitor students' knowledge. These tests reduce time spent on curricula, but serve as diagnostic tools and accountability instruments, alerting teachers and administrators to low-performing student populations in need of additional services or reform. The diagnostic function of testing has merit, but there is a second benefit of testing that is often overlooked: Testing enhances memory for the tested material. Taking advantage of the memorial benefits of tests, and integrating testing into the curriculum rather than as an event that *follows* instruction and learning, has the potential to increase the efficiency and utility of school testing practices if this finding were better understood.

A survey of naive undergraduates supports the claim that tests are viewed principally as assessments in the United States. Kornell and Bjork (2007) asked undergraduates whether they tested themselves when they were studying, and if so, why. Whereas most students did report testing themselves (91%), most stated that they did so to "to figure out how well I have learned the information I'm studying." Only 18% described their testing as a learning event (Kornell & Bjork, p. 222).

## Tests as Learning Events

Research suggests that testing information that has already been studied not only provides a measure of learners' knowledge, tests also become learning events in their own right. Indeed, testing has often been shown to be more effective than further study in encouraging retention of tested information (e.g., Bjork, 1988; Carrier & Pashler, 1992; Gates, 1917; Glover, 1989; Hogan & Kintsch, 1971; Izawa, 1970; McDaniel, Roediger, & McDermott, 2007; Roediger & Karpicke, 2006a, 2006b; Rothkopf, 1966; Tulving, 1967; Whitten & Bjork, 1977; for a review, see Richland, Bjork, & Linn, 2007). Researchers studying the cognitive underpinnings of testing have argued that testing should be considered a strategy for knowledge acquisition above and beyond its utility as a measure of current knowledge.

Testing as an instrument serving larger instructional goals has traditionally been seen to have a limitation, however: The benefits of testing are most pronounced for test items that were answered correctly (Butler & Roediger, 2007; Karpicke & Roediger, 2007;

Lindsey E. Richland and Liche Sean Kao, Department of Education, University of California, Irvine; Nate Kornell, Department of Psychology, Williams College, Los Angeles.

The Office of Naval Research Grant N000140810186 partially supported the experiments reported herein. This material is also based on work supported by the National Science Foundation under Grant 0757646. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We thank H. L. Roediger for insightful comments on the article, and Keara Osborne for invaluable assistance in running participants. Experiments 1, 2, and 4 were previously published in the proceedings of the Cognitive Science Society (Richland, Kao, & Kornell, 2008).

Correspondence concerning this article should be addressed to Lindsey E. Richland, Department of Education, University of California, 2001 Berkeley Place, Irvine, CA 92697. E-mail: l.e.richland@uci.edu

Leeming, 2002; Roediger & Karpicke, 2006a, 2006b). Generally, items not retrieved correctly when tested see minimal, if any, benefit of testing when compared with being allowed additional study time (for exceptions, see Izawa, 1970; Kane & Anderson, 1978; Kornell, Hays, & Bjork, in press). Unsuccessful tests may even have negative consequences. Proponents of errorless learning (e.g., Guthrie, 1952; Skinner, 1958; Terrace, 1963) suggest that failing to answer a question or answering incorrectly makes future errors more likely. Furthermore, being measured alters knowledge representations, and sometimes questioning can lead to memory distortions (see Davis & Loftus, 2007; Roediger & Marsh, 2005). Thus, testing has the potential to distort knowledge, particularly for any items not recalled correctly.

Providing detailed feedback after a test can ameliorate some of these challenges (e.g., Butler, Karpicke, & Roediger, 2007; Kang, McDermott, & Roediger, 2007; Metcalfe & Kornell, 2007; Pashler, Cepeda, Wixted, & Rohrer, 2005), but this type of feedback is burdensome and often not feasible. This is particularly true in standardized testing, when feedback is rarely individualized by question and is often available to students and teachers only after a substantial delay. Thus, for the lowest performing students, who are No Child Left Behind's foremost priority, testing—in particular, failed tests—may have little value (or worse).

## Can Failed Tests Improve Future learning?

The current research posits that the benefits of testing may extend to items that are not answered correctly on the test, and that failure to answer test questions should not be equated with a failure to learn. Rather, five experiments were conducted to evaluate the impact of restructuring the testing environment to actually incur *more* failed tests. Specifically, we evaluated the benefits of testing novel science instructional content *before* learning. Thus, the likelihood of failed tests was high, but we were able to extend our theory of testing to better understand whether trying and failing on test questions actually improved learners' longer term retention of subsequently presented information.

Pretests are regularly used as assessments in pre–posttest design studies with the expectation that they do not affect learning. There were some reasons, however, to expect that pretesting could enhance learning. Many studies have demonstrated benefits of pretraining activities such as advanced organizers (see Huntley & Davies, 1976; Mayer, 1979), outlines (e.g., Snapp & Glover, 1990), and statements to activate learners' prior knowledge schemas (e.g., Bransford & Johnson, 1972). Test questions have also been studied as pretraining activities, beginning with early experiments on the effects of integrating adjunct questions into text passages (e.g., Anderson & Biddle, 1975; Huntley & Davies, 1976; Pressley, Tanenbaum, McDaniel, & Wood, 1990; Rothkopf, 1966; Sagerman & Mayer, 1987). Adjunct questions interwoven into texts, both before and after the target information had been provided, showed improved retention of information asked about in the question and, less reliably, information not asked about (see Anderson & Biddle, 1975; Mayer, 2008; Rickards, 1976).

This basic pattern has held for both direct questions with basic text materials and more complex learning environments with higher level questions. For example, using "deep-level-reasoning questions" to introduce and frame interactions with an automated tutoring system, Autotutor, can greatly affect learning (e.g., Craig, Gholson, Ventura, Graesser, & the Tutoring Research Group, 2000; Craig, Sullins, Witherspoon, & Gholson, 2006; Gholson & Craig, 2006). In some circumstances, integrating these questions into instructional content can make noninteractive instruction as effective as an interactive tutor (VanLehn et al., 2007). Related research demonstrates that training "self-questioning" improves critical thinking and learners' ability to construct knowledge from forthcoming instruction (see King, 1992, 1994).

The early studies on adjunct questions, and the more recent studies with more inferential, higher level questions, did not attempt to contrast failures at the time of testing with successes. Rather, the most common interpretation of the questions' effects on later retention rested on their impact on readers' intentional learning behaviors. Rothkopf (1965, 1966, 1982) coined the term *mathemagenic behaviors* to explain the intentional learning behaviors of readers that are alterable by the instructional activities they encounter. For example, Rothkopf and Bisbicos (1967) found that asking participants questions in which the answers were numbers led to better retention of all numerical information in the text, possibly because participants were able to direct their attention to the type of information that was important to learn given the test they would take.

Direct tests of attention, based on measures of reading time and reaction time to a secondary task, demonstrate that people pay greater attention to reading a text when adjunct questions are interwoven (Reynolds & Anderson, 1982; Reynolds, Standiford, & Anderson, 1979). A practice guide published by the Institute of Education Sciences (an institute within the U.S. Department of Education) reviewed recent research with a similar conclusion, making the instructional recommendation: "We recommend . . . using 'prequestions' to activate prior knowledge and focus students' attention on the material that will be presented in class" (Pashler et al., 2007, p. 30).

In addition to affecting learners' attention and intentional learning behaviors, pretesting may provide a direct impact on memory. The cognitive benefits of testing *after* studying are well established to persist even when there is no opportunity to restudy information (e.g., Hogan & Kintsch, 1971; Roediger & Karpicke, 2006b), which rules out the possibility that those benefits are explainable by attention during text processing. We thus investigated whether there was a similar cognitive benefit for pretesting above and beyond the effect of drawing learners' attention to testable information.

Unlike most previous studies, Pressley et al. (1990) did distinguish the effects of attention direction from the effects of testing itself using pretest questions. Their participants recalled more when they had been asked about the passage before reading it than when they had been presented with the same pretest questions, but had not been asked to try to answer them, before reading the passage (instead, participants were asked whether or not the questions were well written). Because the questions had dichotomous answers, however, participants were frequently able to answer correctly during the pretest.

The current experiments followed a similar study premise, but sought to test more directly whether unsuccessful retrieval attempts enhance retention of tested content beyond directing attention during study. Therefore, in the current study, the prequestions required participants to produce nouns or descriptive statements that they were unlikely to be able to answer on the basis of prior

knowledge (e.g., "What is total colorblindness caused by brain damage called?"). This allowed us to isolate and examine the effects of unsuccessful retrieval attempts.

The questions tested knowledge for exact information presented in the text, rather than knowledge that would require inferential or higher level thinking. Such questions are effective (e.g., Marsh, Roediger, Bjork, & Bjork, 2007; Rickards, 1976; Rickards & Hatcher, 1977–1978; Watts & Anderson, 1971; Yost, Avila, & Vexler, 1977) but could not be used in this study. They would have prevented us from adequately controlling for the fact that test questions draw learners' attention to testable content. Rather, we wanted to be able to create a no-pretest control condition in which we could draw participants' attention to the same information asked about in the test questions.

Typographical cuing (e.g., underlining or bolding; for reviews, see Glynn, Britton, & Tillman, 1985; Waller, 1991) is effective at drawing attention to cued items, sometimes to the exclusion of uncued items (Glynn & DiVesta, 1979). By typographically cuing participants to the aspects of the passages that would be tested, we expected to draw their attention to the key content that needed to be learned. This would allow us to distinguish between the effects of attention direction and any additional benefits of unsuccessful retrieval attempts.

## The Present Experiments

We report four experiments that examined the learning effects of pretesting, beyond directing attention to testable information, when the questions were answered incorrectly. Theoretically, we sought to analyze the effects of attempting (but failing) to retrieve or generate test answers from memory, as distinct from participants' use of more directed search strategies while reading the text. In a fifth experiment, we further distinguished between attempting to retrieve answers to test questions and other deep processing of the pretest questions.

In all experiments, participants were asked to read a scientific text about vision in an unstructured reading situation, akin to how a learner might study a textbook. In the first experiment, participants were either tested prior to learning or they were given additional time to study. In Experiments 2 through 5, variations on the same procedure were used to isolate the effect of attempting to derive an answer to a question from the more mundane effect of directing attention by preexposing questions. In Experiment 2, all tested sentences were italicized in the studied text; in Experiment 3, the keyword from each tested sentence was bolded. Experiment 4 used bolded text and assessed the impact of testing versus extended study after a 1-week delay. Experiment 5 sought to differentiate between *reading* potential test questions and attempting to *answer* test questions before studying. Similar to Pressley et al. (1990), we manipulated whether participants memorized the pretest questions versus produced an answer to the same questions.

## Experiment 1

We predicted that testing before study would enhance future recall, in spite of learners' failure to provide successful answers to the test questions.

### *Method*

#### *Participants*

Participants in this study were 63 undergraduates who were given extra course credit for participating.

#### *Materials*

Study materials were selected from Sacks (1995). A two-page text was developed on the basis of an essay about a patient with cerebral achromatopsia (colorblindness caused by brain damage). This text was selected because of its rich scientific content and engaging narrative. The reading level was deemed appropriate for undergraduates, and Sacks's book is assigned in undergraduate coursework. To protect against the possibility that participants had read the passage in coursework, participants were asked whether they had read the passage previously, in which case they would have been excluded. None were excluded for this reason. The length of the story was designed to ensure that participants were not under time pressure and had time to return to sections if they desired to do so.

Some of the text described Sacks's patient suffering from cerebral achromatopsia, as in the following sample:

> I am a rather successful artist just past 65 years of age. On January 2nd of this year I was driving my car and was hit by a small truck on the passenger side of my vehicle. When visiting the emergency room of a local hospital, I was told I had a concussion. . . . I have visited ophthalmologists who know nothing of this color-blind business. I have visited neurologists, to no avail. Under hypnosis I still can't distinguish colors. I have been involved in all kinds of tests. You name it. My brown dog is dark gray. Tomato juice is black. Color TV is a hodgepodge.

Other parts of the text were selected from the more scientific treatment of the disorder, as in the following sample:

> Colorblindness, as ordinarily understood, is something one is born with—a difficulty distinguishing red and green, or other colors, or (extremely rarely) an inability to see any colors at all, due to defects in color responding cells, the cones of the retina. Total colorblindness caused by brain damage, so-called cerebral achromatopsia, though described more than three centuries ago, remains a rare and important condition. It has intrigued neurologists because, like all neural dissolutions and destructions, it can reveal to us the mechanisms of neural construction, specifically, here, how the brain "sees" (or makes) color. (Sacks, 1995, pp. 3–4)

Within the reading packet, 10 sentences were identified as testable items. Test questions were constructed on the basis of these 10 sentences. Two counterbalanced pretests were constructed such that each contained questions about 5 of the selected sentences. Questions were written as fill-in-the-blank or short free-response items (e.g., "What is total color blindness caused by brain damage called?" and "How does Mr. I distinguish red and green traffic lights?"). They addressed facts presented in the text, either general scientific facts or information about the specific patient. See Appendix A for all questions.

A final test included all 10 of the testable items in randomized order. Thus, for participants in the test condition, 5 of the final test questions had been pretested during Time 1 (tested) and 5 had not

been tested previously (untested). Questions from the two pretest versions were always interspersed on the final test. All questions were new for participants in the extended study condition.

### Procedure

The experiment was conducted in a group setting. Participants were randomly assigned to an *extended study* condition ($n = 27$) or a *test and study* condition ($n = 36$). We conducted this experiment in a lecture class setting, and assigned participants on the basis of seating. The lecture space had separate seating areas and participants were assigned on the basis of those. This was done to ensure that each section followed the appropriate timing, but we did not have tight control over cell size.

*Learning phase.* Participants in the test and study condition were given one of the two counterbalanced pretests and allowed 2 min to answer the questions. They were instructed to provide an answer to all five questions, regardless of whether they knew the answer. At the end of 2 min, the pretests were collected, and participants given the text passage and told to study it for 8 min. They were instructed to read the passage through in its entirety at least once.

Participants in the extended study condition were given 10 min to study the passage—the same total time that participants in the test and study condition spent in testing and study of the material. They were given the same reading instructions.

*Final test.* The text passages were collected after the timed study periods were completed. Participants were then immediately administered the Time 2 test, which consisted of 10 questions. The test was untimed to ensure that time pressure did not affect performance.

## Results

In the test and study condition, on the initial test that preceded the presentation of the passage, participants answered 5% of the questions correctly. Any items answered correctly on the Time 1 pretest were removed from the following analyses of Time 2 test scores on a participant-by-participant basis. Most participants gave an answer for all questions, often providing answers that were incorrect yet appropriate (e.g., writing the name of a scientist in a question referring to Isaac Newton).

An independent samples *t* test examined the effects of testing by comparing mean posttest percentage correct for tested items in the test and study condition with the overall mean score in the extended study condition. As shown in Figure 1, testing resulted in better posttest performance ($M = 75\%$, $SE = 3.2$) than did the provision of extra time to study the same material ($M = 56\%$, $SE = 2.7$), $t(61) = 4.26$, $p < .0001$, $d = 1.1$.

Examining performance within the test and study condition only, tested items ($M = 75\%$, $SE = 3.2$) were recalled on the final test significantly more often than untested items ($M = 50\%$, $SE = 3.4$), $t(35) = 5.03$, $p < .0001$, $d = 1.7$, in spite of the fact that the analyses excluded any items that participants recalled correctly on the pretest. The benefit of testing did not spread to untested items, but neither did it hurt. There was not a significant difference between accuracy on the untested items in the test and study condition and in the extended study condition, $t(61) = 1.3$, $p = .20$.

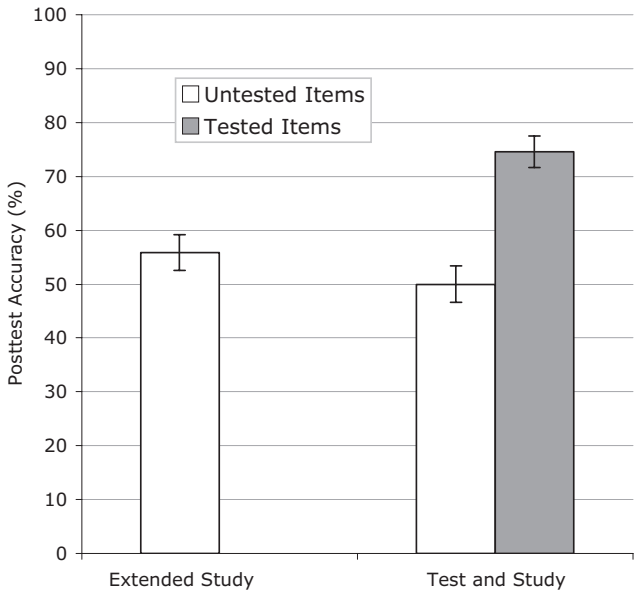


*Figure 1.* Experiment 1: Performance on a final test across conditions when studying an unmarked text.

## Discussion

Experiment 1 revealed that failed tests *can* enhance learning for educational content. Although participants largely failed on the initial test (answering 95% of the questions incorrectly), the effect of those failures was to increase retention of studied content when compared with an extended opportunity to study the materials without being pretested.

The explanation for the benefit of unsuccessful tests is not yet clear. One possibility is that the test directed learners' attention to the key, testable points in the passage. Alternatively, attempting to retrieve an answer to the test problem may have provided an additional benefit above and beyond the impact of attention direction. Experiment 2 used the same procedure as Experiment 1, but all testable sentences were italicized to equalize participants' attention to key concepts in the text. We reasoned that under such conditions, allocation of attention would not differ meaningfully between conditions; therefore, differences in learning would be attributable to the impact of retrieval attempts during the pretest.

# Experiment 2

We predicted that pretesting followed by study would enhance future recall more than the provision of extended time to study an instructional text, even when differences in attention direction were minimized by italicizing key sentences in the text in both conditions.

## Method

### Participants

The participants were 61 undergraduates (mean age = 21 years, 44 women and 17 men) who were given extra course credit for participating. Participants were sampled from an upper division

psychology course on human stress. Data from 2 participants were excluded from analyses because of a failure to respond to final test questions.

### Materials

The study materials were the same text and testable sentences as used in Experiment 1. The key difference was that within the reading packets, the 10 testable sentences were italicized. Italicizing was considered a way to ensure that all participants were equally alerted to what was deemed to be important information in the same way that many textbooks emphasize key elements of a chapter. Participants in both conditions read the same italicized text. For example, see the following text paragraph:

> The history of our knowledge about the brain's ability to represent color has followed a complex and zigzag course. *Newton, in his famous prism experiment in 1666, showed that white light was composite—could be decomposed into, and recomposed by, all the colors of the spectrum.* The rays that were bent most ("the most refrangible") were seen as violet, the least refrangible as red, with the rest of the spectrum in between. (Sacks, 1995, p. 18)

### Procedure

The procedure was exactly the same as the procedure in Experiment 1. Participants were tested in a group setting and were randomly assigned to the extended study condition ($n = 26$) or to the test and study condition ($n = 33$). Participants were not given any specific instruction regarding the text italics.

## Results

In the test and study condition, participants answered 22% of questions on the initial pretest correctly. Correct answers were distributed across test problems. The population of participants in this experiment seems to have had a higher level of relevant background knowledge on pretest items than in Experiment 1, perhaps because they were sampled from a higher level psychology course, but as in Experiment 1, any items answered correctly at Time 1 were removed from the following analyses on a participant-by-participant basis. If anything, this led to inflation in participants' scores in the untested conditions, counter to our hypothesis.

The data revealed benefits for testing over the provision of extra time for studying the same material. As Figure 2 shows, recall of tested and italicized items in the test and study condition ($M = 71\%$, $SE = 5.6$) was significantly greater than recall of italicized-only items in the extended study condition ($M = 54\%$, $SE = 3.7$), $t(57) = 2.3$, $p < .022$, $d = 0.61$.

Examining performance within the test and study condition, tested items were recalled on the final test ($M = 71\%$, $SE = 5.6$) significantly more often than untested, italicized-only items ($M = 53\%$, $SE = 4.3$), $t(32) = 3.27$, $p < .003$, $d = 0.63$, in spite of the fact that the analyses excluded items that participants recalled correctly on the pretest. Testing did not appear to negatively affect the untested items; there was not a significant difference between accuracy on the italicized-only items in the test and study condition and the italicized-only items in the extended study condition, $t(57) = 0.15$, $p = .88$.

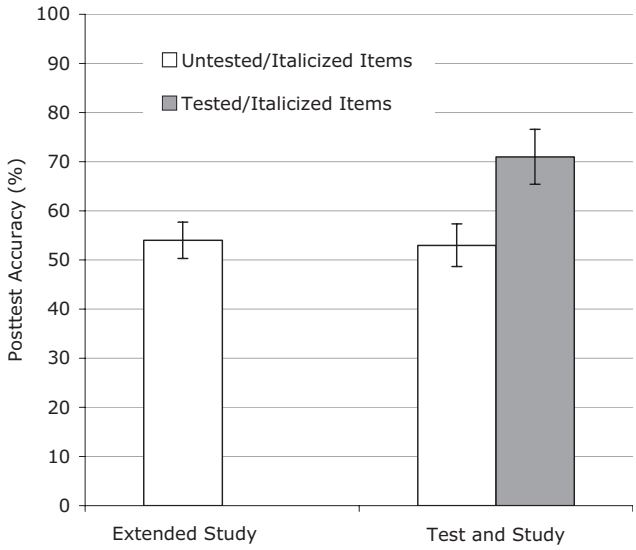


*Figure 2.* Experiment 2: Performance on a final test across conditions when studying text with italicized key sentences.

## Discussion

The results of Experiment 2 replicated the results from Experiment 1, and again suggest that the testing effect can and should be extended to failed tests. Testing items created more potent learning opportunities than extended study of the same items, even when the key information in both conditions was italicized, equalizing attention direction. Thus, testing appears to provide a unique benefit above and beyond directing learners' attention to content that has a high probability of being tested later.

In textbooks, italicized sentences are less common than bolded keywords, which are ubiquitous. It remains possible that participants in Experiment 2 were unfamiliar with the meaning of italics within text, and thus differences in attention were not minimized. To rule out that possibility, Experiment 3 used the same procedure as Experiment 2, but bolded keywords were used instead of italicized sentences because we expected that bolding might act as a stronger (and more realistic) attention prompt. Experiment 3 thus examined the impact of testing when compared with extended opportunities to study text in which the key test items were bolded.

# Experiment 3

We predicted, similar to Experiment 2, that testing before reading would enhance future recall above and beyond the impact of extended study time. Instead of presenting key sentences in italics, keywords were presented in bold.

## Method

### Participants

Participants in this study were 64 undergraduates (44 women, 17 men, 3 unstated) who were given extra credit in their courses for participating. Participants' average age was 22 years.

### Materials

The test materials were exactly the same as those used in Experiments 1 and 2. The study materials were exactly the same as those used in Experiments 1 and 2, with the exception of the treatment of the 10 testable sentences. Within the reading packet, one word was bolded from each of the sentences that had been italicized and tested in Experiment 2. The bolded word was the answer to the fill-in-the-blank or short-answer questions used in the tests. An example of a paragraph with bolded words follows:

> Colorblindness, as ordinarily understood, is something one is born with—a difficulty distinguishing red and green, or other colors, or (extremely rarely) an inability to see any colors at all, due to defects in color responding cells, the **cones** of the retina. Total color blindness caused by brain damage, so-called **cerebral achromatopsia**, though described more than three centuries ago, remains a rare and important condition. (Sacks, 1995, pp. 3–4)

### Procedure

The procedure was exactly the same as the procedure in Experiments 1 and 2. The experiment was conduced in a group setting. Participants were randomly assigned to the extended study condition ($n = 33$) or the test and study condition ($n = 31$). No specific instructions were given regarding the bolded text.

### Results

In the test and study condition, on the initial test that preceded the presentation of the passage, participants answered 21% of the questions correctly. Two pretest items about vision were answered correctly at unexpectedly high rates, something that had not occurred in the previous experiments, so these questions were removed from all further analyses of posttest data for this experiment in both conditions. Excluding those questions led to a pretest average performance level of 11%. Any other items answered correctly at Time 1 were removed from the following analyses on a participant-by-participant basis.

As shown in Figure 3, tested and bolded items in the test and study condition were recalled significantly more often on the final test ($M = 82\%$, $SE = 3.8$) than were bolded-only items in the extended study condition ($M = 64\%$, $SE = 4.0$), $t(62) = 3.3$, $p < .002$, $d = 0.84$, revealing a benefit for testing over extra time spent studying the same material. Even when keywords were bolded in both conditions, pretesting led to higher retention of bolded and tested items than did extended study.

Within the test and study condition, there was a numerical advantage for tested and bolded items ($M = 82\%$, $SE = 3.8$) over items that were bolded but not tested ($M = 77\%$, $SE = 3.0$), but unlike in Experiments 1 and 2, the difference was not significant, $t(30) = 1.4$, $p = .17$. This lack of difference may indicate that even untested items benefited from testing. Indeed, untested items in the test and study condition were recalled at a higher rate than items in the extended study condition, a difference that approached significance, $t(62) = 1.9$, $p = .062$, $d = 0.48$. Although this finding was not reliable across all studies reported herein, it is consistent with the early arguments that testing before learning affects readers' intentional learning practices. At minimum, these data suggest that testing did not hurt recall of untested items when keywords were bolded.

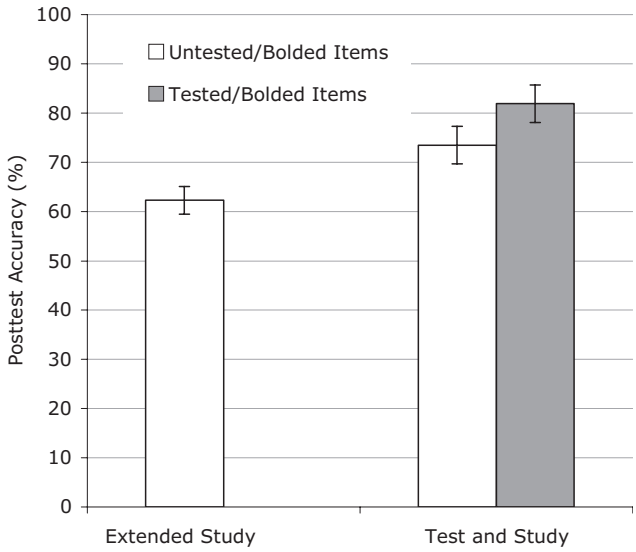


*Figure 3.* Experiment 3: Performance on a final test across conditions when studying text with bolded keywords.

### Discussion

Experiment 3 demonstrated that unsuccessful tests can enhance learning for new educational content, replicating and extending the findings from Experiments 1 and 2. Testing items before learning was a more potent learning opportunity than the provision of extended study time, even when keywords were bolded in the text and only items that participants failed to answer on the initial test were included in the analyses. Once again, these results suggest that testing provides a unique benefit above and beyond serving to direct learners' attention to materials that might be tested at a later point. The results of Experiment 3 also suggest that testing some items may additionally benefit learning for untested items.

## Experiment 4

In the first three experiments, the effects of pretesting were measured on an immediate test. Previous research has shown, in the context of successful tests, that the size of the testing effect grows as the delay between study and a final memory test increases because tested items are forgotten more slowly than items that have not been tested (Hogan & Kintsch, 1971; Roediger & Karpicke, 2006b). In Experiment 4, to investigate the effect of delaying the final test for items that have been tested unsuccessfully, we examined learning after a 1-week delay. Doing so was also a way to connect the findings with the goals of education, which involve improving long-term learning. There was also a second change to Experiment 4. To better distinguish between the effects of bolding and testing, we manipulated bolding within subjects. Testing versus extended study remained a between-subjects manipulation.

We predicted that the results would be similar to the results of the previous experiments—that is, that final test performance at a

delay would be higher for items that were pretested in the test and study condition than for items that were bolded in the extended study condition, even if the retrieval attempts on the pretest were unsuccessful. We also predicted that bolding would benefit retention relative to nonbolded items in the extended study condition, but that testing would be more advantageous than bolding.

## Method

### Participants

Participants in this study were 158 undergraduates (137 women, 15 men, 6 not stated; mean age = 20 years) who were given extra course credit for participating.

### Materials

The study materials were exactly the same as those used in Experiment 3 with one exception: Rather than emphasizing all 10 key concepts, as in Experiments 2 and 3, only 5 items were bolded. In the test and study condition, the 5 items tested on the Time 1 pretest were the same items that were bolded. In the extended study condition, 5 corresponding items were selected to be bolded. These were matched to the items tested in the counterbalanced test conditions. Thus, for a given participant, of the 10 items tested on the posttest, 5 had been emphasized during initial study (by being bolded in the extended study condition, or by being tested and bolded in the test and study condition), and 5 items had not. Tested and bolded items were counterbalanced across participants. This manipulation allowed us to make separate estimates of the effects of bolding and the effects of testing.

In addition, two questions that had received relatively high accuracy rates on the pretest were rewritten. See Appendix B for replacement questions.

### Procedure

The learning phase of the experiment was identical to the learning phase of Experiments 1–3, except that, to control the timing of the final test, participants were tested individually. After completing the first session, participants were asked to return 1 week later at the same time of day. When they returned, the final test was administered. The test procedure was the same as the tests in the previous experiments. Participants were randomly assigned to the extended study condition ($n$ = 79) or the test and study condition ($n$ = 79).

## Results

In the test and study condition, on the initial test that preceded the presentation of the passage, participants answered 10% of the questions correctly. Items answered correctly on the pretest were removed from the analyses on a participant-by-participant basis.

The data were analyzed differently from those in Experiments 1–3 because posttest performance for both conditions could be separated into bolded and nonbolded items. Thus, there was a within-subjects manipulation of bolding and a between-subjects manipulation of testing. Because testing and bolding were manipulated together in the test condition (items were tested and bolded or untested and unbolded), this is not a full factorial design and effects were analyzed using one-tailed $t$ tests. The effects of testing were examined by holding bolding constant between the testing and extended study conditions (bolded and tested vs. bolded). The effects of bolding were studied in the extended study condition (bolded vs. unbolded).

The results are displayed in Figure 4. There was a significant pretesting effect: Bolded and tested items in the test and study condition were recalled better ($M$ = 55%, $SE$ = 2.0) than bolded-only items presented for longer study time in the extended study condition ($M$ = 45%, $SE$ = 3.0), $t(156) = 2.8$, $p < .0025$, $d = 0.45$. When the test and study condition was examined separately, tested and bolded items were recalled at a distinctly higher rate ($M$ = 55%, $SE$ = 0.30) than untested and unbolded items ($M$ = 42%, $SE$ = 3.0), $t(78) = 3.3$, $p < .001$, $d = 0.75$. There was also a smaller difference between bolded items and unbolded when the extended study condition was examined separately, $t(78) = 1.9$, $p < .04$, $d = 0.43$ ($M$ = 45%, $SE$ = 2.2, and $M$ = 39%, $SE$ = 2.6, respectively), revealing that bolding was an effective typographical tool for directing attention. There were no significant differences between the test and study condition and the extended study condition on untested and unbolded items, $t(156) = 0.58$, $p = .28$ ($M$ = 42%, $SE$ = 3.0, and $M$ = 39%, $SE$ = 3.0, respectively).

## Discussion

Experiment 4 demonstrated that failed tests can affect learning for educational content even after a 1-week delay, extending the findings from Experiments 1–3. Once again, the results suggest that testing provides a unique benefit above and beyond serving to direct learners' attention to materials that might be tested at a later point. Indeed, directing attention by bolding items provided a minimal benefit in the extended study condition, whereas bolding

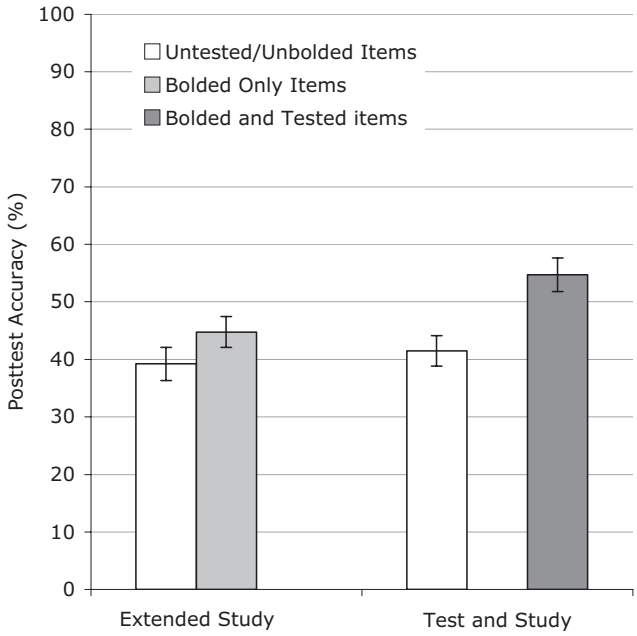


*Figure 4.* Experiment 4: One-week delayed performance on a final test across conditions when keyword bolding was manipulated within subjects.

accompanied by testing significantly enhanced learning in the test and study condition.

Although previous results have shown that testing effects sometime increase as the delay between study and final test increases (e.g., Roediger & Karpicke, 2006a), the sizes of the effects in Experiment 4 were comparable to the sizes of the effects in Experiment 3. This finding suggests that, unlike successful tests, unsuccessful tests may not slow the rate of forgetting, although further evidence would be needed to support that hypothesis.

## Experiment 5

In Experiment 5, we investigated why unsuccessful tests enhance learning. Specifically, we sought to determine whether the pretesting effects that we had identified could be attributed to attempting (albeit unsuccessfully) to *answer* test questions versus simply seeing potential test questions before beginning to study. Providing test questions, even if participants do not have to answer them, could have similar results to pretesting, and thus explain the pretesting effects in Experiments 1–4 without reference to the direct benefits of tests. Two explanations for those results could include that (a) test questions may provide an organizational framework that indirectly affects retention by guiding future learning, and (b) allowing participants to read test questions may induce deep processing more effectively than does merely reading the passage.

Providing potential posttest questions to readers before reading the passage may serve as a guide for readers' interactions with the forthcoming text, either as an organizational framework to better structure causal structure and knowledge interpretations (e.g., Craig et al., 2000; Pashler et al., 2007) or by affecting learners' looking behaviors (Rothkopf, 1965, 1982). If this is the case, simply reading the questions before studying may be as effective as attempting to answer them—perhaps more so if answering questions incorrectly leads to retention for those incorrect answers (Marsh et al., 2007; Roediger & Marsh, 2005).

Alternatively, the levels of processing theory have been posed to explain the benefits of testing in general and may apply to both successful and unsuccessful tests. For example, Kane and Anderson (1978) found that asking participants to fill in the last word in a sentence helped them remember the correct last word, even when most of the words the participants filled in were incorrect. These authors hypothesized that testing resulted in a deeper level of processing than simply reading, which served to organize sentence information in participants' minds and make it recallable on a later test.

Ghatala (1981) provided further support for the notion that testing benefits learners mainly because it induces deep processing. Participants were asked to fill in the missing last word of a sentence; unlike Kane and Anderson's (1978) materials, the missing word was obvious and participants usually retrieved it successfully. Ghatala compared the retrieval condition to a condition in which participants did not retrieve, but were asked to do a task that induced deep processing. Ghatala found that "the operations involved in generating information from semantic memory have no special mnemonic value beyond inducing optimal processing of the material" (p. 443). For questions in which the missing word was obvious, retrieving the key word was no better than other deep processing of the sentence.

It is interesting that a follow-up study indicated that testing might produce an additional benefit over deep processing when the missing word was not obvious (Ghatala, 1983). Ghatala interpreted these combined results as suggesting that attempting to retrieve did not by itself provide a direct mnemonic advantage, but could indirectly improve learning by strengthening memory for the sentence's organizational structure. This echoes the benefits of providing "prequestions" or other advanced learning techniques to organize forthcoming knowledge.

For a different interpretation, one may consider the Pressley et al. (1990) findings, reviewed above, which showed benefits of pretesting greater than the benefits of viewing test questions before learning. When taken together with the Ghatala (1983) data, these findings suggest that failed tests may affect retention both directly and indirectly.

Experiment 5 investigated whether the benefits of unsuccessful tests result from the active attempt to recall key information from memory versus simply more active processing of the test sentence as an organizational structure. We compared the two conditions from the previous four experiments (i.e., extended study and test and study), as well as a third condition. In the third condition—the question learning condition—participants were asked, before reading the passage, to memorize the test questions without attempting to answer them (similar to the procedure used by Pressley et al., 1990). We expected that trying to memorize the questions would induce a relatively deep level of processing as participants focused on integrating the semantic structure of test sentences. Thus, any advantage of processing the test questions as organizational structures should be comparable across the two prequestion conditions.

This procedure led to two conflicting predictions. On the basis of Ghatala's (1981, 1983) results and interpretation, we anticipated that testing and question learning might have equivalent effects because both induce deep processing and support learning for test sentences as organizational structures. On the basis of the hypothesis that attempting to retrieve is more effective than deep processing alone, however, and in agreement with Pressley et al. (1990), we predicted that pretesting would lead to higher retention rates than would attempting to memorize and reproduce test questions without answering them.

### *Method*

#### *Participants*

The participants in Experiment 5 were 76 undergraduates (64 women, 12 men), with an average age of 20 years, who were given extra credit in their courses for participating. An additional 3 participants were excluded from the analyses for failing to complete the posttest, and 2 were excluded for prior knowledge of the tested passage.

#### *Materials*

The study materials were similar to those used in Experiment 4; for any given participant, half of the key concepts in the text were emphasized. We returned to italicizing multiple words (as in Experiment 2) rather than bolding single words (as in Experiments 3 and 4) because doing so provides more information to the learner about exactly what information to focus on. Italicizing might also

be an effective way to focus participants' attention because of its novelty. Instead of italicizing whole sentences, as in Experiment 2, we italicized key phrases within the sentences to make the direction of attention more precise (e.g., "*Total color blindness caused by brain damage, so-called cerebral achromatopsia*, though described more than three centuries ago, remains a rare and important condition").

The test questions used in the pre- and posttests were modified versions of the tests used in Experiments 3 and 4. Because participants would be memorizing the questions, we wanted to minimize difference in form, length, and number of untested facts included in the question text. All questions were rewritten to fill-in-the-blank format and longer questions were simplified and shortened. See Appendix C for modified questions.

### *Procedure*

The procedure was the same as Experiments 1, 2, and 3, except that there was a third between-participants condition, in which a new set of instructions was given during the Time 1 pretest. Participants were tested individually and were randomly assigned to one of three conditions: extended study ($n = 26$), test and study ($n = 24$), and question learning and study ($n = 26$).

The procedures for the extended study and test and study conditions were the same as Experiments 1–4. In the question learning and study condition, participants were given the same counterbalanced Time 1 tests as were used in the test and study condition. Instead of being asked to answer the questions, however, participants were asked to memorize the test questions because, they were told, they would be testing another person on the questions. They were asked to pay careful attention to where the blank fell in the question. This instruction was intended to support the students in learning the question without filling in an answer. After 2 min of studying the questions, participants were given a blank sheet of paper and asked to write down the questions. They were again cautioned to make sure to leave a blank in the correct spot. This procedure step provided an additional level of processing the question.

### *Results*

In the test and study condition, on the initial test that preceded the presentation of the passage, participants answered 6% of the questions correctly. All items answered correctly on the initial test were removed from analyses of posttest performance.

The results are shown in Figure 5. One-tailed planned comparisons first examined the hypothesis that participants in the test and study condition would outperform participants in the question learning and study and extended study conditions on ability to answer posttest questions on italicized keywords. The test and study condition outperformed the question learning and study condition, $t(48) = 2.04$, $p < .02$, $d = 0.59$, which in turn outperformed the extended study condition, $t(50) = 2.02$, $p < .02$, $d = 0.57$ (test and study: $M = 90\%$, $SE = 3.8$; question learning: $M = 78\%$, $SE = 4.2$; extended study: $M = 63\%$, $SE = 5.9$). As expected, the largest difference was between the test and study condition and the extended study condition, $t(48) = 3.7$, $p < .001$, $d = 1.1$.

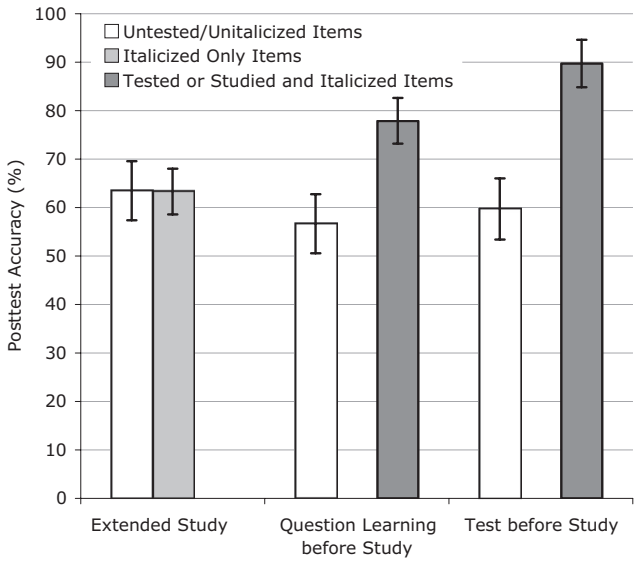


*Figure 5.* Experiment 5: Performance on a final test following varied pretest learning activities and study.

To examine any related differences on untested items, a one-way analysis of variance compared performance across conditions on items that were neither tested nor italicized. There were no differences between conditions on these items, $F(1, 73) = 0.31$, $MSE = 0.096$, $p = .74$.

### *Discussion*

The results of Experiment 5 extend findings from the previous experiments demonstrating the benefit of attempting to retrieve a response to a question, even when the attempt is unsuccessful. Attempting to answer a prequestion was significantly more effective than reading the same question and attempting to memorize it without making an attempt to retrieve the answer. It is even possible that the benefits of the question learning and study condition were attributable to participants attempting to answer some of the questions despite being asked not to—in essence, to the benefit of testing. These data suggest that the benefits of testing extend beyond the benefits of engaging in deep processing. Most important, the data support the hypothesis that unsuccessful tests are useful because of their role as tests, apart from the role prequestions may play in encouraging deep processing or supporting organization of forthcoming knowledge.

## General Discussion

Previous research has demonstrated the memory benefits of successfully answering test questions. The five experiments reported herein provide evidence for the power of tests as learning events even when the tests are unsuccessful. Participants benefited from being tested before studying a passage—a pretesting effect—although they did not answer the test questions correctly on the initial test, as compared with being allowed additional study time. Furthermore, the benefits of pretests persisted after a 1-week delay. Tests can direct participants' attention to the important information, but such attention direction cannot explain the current

findings because the important information was highlighted in all conditions using italics or bolding in Experiments 2–5. Moreover, participants in Experiment 5 learned more after unsuccessfully attempting to answer test questions than they did after attempting to memorize, but not answer, the same questions. These data imply that testing has advantages that exceed the benefits prequestions may have in supporting the organization of knowledge structures or in giving rise to deep processing (see also Kornell et al., in press).

There was little cost to testing; it did not require the provision of additional time on task, and nontested items were not adversely affected when other items were tested in the test condition. The effects on untested items varied between experiments, showing a positive effect of testing in Experiment 3 but no significant differences across the other studies. These data do not seem to reveal a systematic pattern, but the important point is that, on the basis of the present findings, pretesting did not seriously impair retention of untested items, as it has been posited to do previously (see Frase, 1968, 1970; Pashler et al., 2007). We tentatively conclude that pretesting can be employed without significant risk to untested items.

### *Theoretical Implications of Unsuccessful Tests*

The present findings suggest that the testing effect—that is, the finding that more learning occurs during testing than when information is presented without a test—is not solely a result of the benefits of successful attempts to retrieve information from memory. Successful tests may play a powerful role in enhancing memory, but attempting to retrieve information, by itself, enhances future learning.

From a cognitive perspective, there are a number of reasons why unsuccessful tests might enhance future learning. One reason is that retrieval strengthens retrieval routes between the question and the correct answer (e.g., Bjork, 1975, 1988; McDaniel & Masson, 1985). Participants frequently generated appropriate but incorrect answers, which might seem more likely to strengthen dead ends than retrieval routes; however, the function of a failed retrieval attempt may be to weaken or suppress errors, rather than to strengthen them (e.g., Carrier & Pashler, 1992). Alternatively, retrieving appropriate content potentially could have strengthened retrieval pathways to related content, identifying a need for additional information, thus strengthening the route even before the content was provided.

A second potential reason for the benefits of unsuccessful tests is that they can encourage deep processing of the question in a way that merely reading the question does not (Bjork, 1975; Carpenter & DeLosh, 2005; Ghatala, 1981, 1983; Kane & Anderson, 1978). To retrieve an appropriate answer to a question, a learner may attempt to imagine or creatively search for potential solutions. For example, even if a learner cannot think of the correct answer to a question such as "How does Mr. I distinguish red and green traffic lights?" the question may prompt the learner to picture a traffic light, think about approaching a traffic light while driving, consider what color blindness is like, what sorts of mishaps one might encounter, and so on. Even when these thoughts do not produce a correct answer, they may create a fertile ground for later encoding of the answer when it is eventually provided, and therefore may produce benefits similar to the effects of deep processing of the answer (e.g., Craik & Lockhart, 1972). Under some circumstances, an unsuccessful retrieval attempt might, by this logic, even result in more learning than a fast, successful retrieval attempt.

We examined the effects of deep processing of the question, in Experiment 5, by presenting participants with pretest questions, but asking them either to try to answer the question or to try to memorize the question. Both instructions were designed to induce deep processing of the semantic meaning of the question. Ghatala (1981) posited that, at least with respect to successful tests, the testing effect was attributable to the benefits of deep processing and internalizing the organizational structure of sentences. The present results suggest that testing was more beneficial than deep processing of the sentence. If participants in the test and study condition in Experiment 5 engaged in the type of thought processes described above (e.g., thought about approaching a traffic light while driving or other information outside of the strict confines of the question), however, testing may have resulted in deeper, more complex levels of processing than question memorization.

Thus, the nature of the processing learners perform during a prelearning activity may be more crucial than the amount of processing performed. Carpenter and DeLosh (2005) found, for tests administered after learning and before a final knowledge assessment, that the degree of elaborative processing required during testing was predictive of final test performance, regardless of the final test format. Free-response tests, which require the most elaborative processing, led to the highest overall retention, whereas recognition and cued recall produced smaller benefits. Similarly, in Experiment 5, attempting to retrieve an answer to the pretest questions could have produced qualitatively more elaborative processing than attempting to learn the test question as an organizational structure, even if the *amount* of processing in the two conditions was similar.

### *Applications to Educational Practice*

Even if tests are not answered successfully, they have the potential to improve future learning, as measured by both immediate and delayed performance measures. This finding suggests that using tests as learning events in educational settings could have lasting benefits for learners' content acquisition, and that tests should be considered a potent learning opportunity, rather than simply as an assessment measure.

According to Bransford and Schwartz (1999), the quality and cognitive impact of a learning event can be measured, in part, by the impact of the learning event on future knowledge acquisition. Bransford and Schwartz emphasize the importance of "preparation for future learning" (p. 8) as a measure of transfer. The current experiments show that one way to prepare learners for future knowledge acquisition is to ask them to answer test questions before studying, even if they are unsuccessful in their attempts.

Although feedback on tests is known to aid learning (e.g., Kang et al., 2007), our data suggest that instruction following testing need not be individualized to learner errors. Rather, instruction that appropriately draws attention to key content may build on the previous cognitive acts performed when attempting to answer a test question. This implies that standardized tests, or other test situations where it is difficult to provide timely item-by-item feedback, could still provide learning benefits for successful and

unsuccessful test takers, as long as those test takers are given an opportunity to learn the information on which they were previously tested. Such a goal, however, would rely on close alignment between the test and subsequent learning opportunities. Standardized tests are not yet so closely aligned with curriculum, although such alignment could potentially enhance the usefulness of both the testing and the instruction that followed it.

At a practical level, pretests could be relatively feasible to implement in classrooms. Although research is necessary to clarify whether the same benefits apply across diverse texts and question types, many teachers already do some pretesting, and teachers could fairly easily make use of test bank or end-of-chapter questions that textbooks almost universally provide. Textbook questions often correspond to bolded keywords in the text, leading to similarities to the current experimental manipulations. As one usage that would be quite similar to the current experiments, pretests might help teachers ensure that students memorized key basic facts for a unit, freeing them to spend more subsequent time on more conceptual and inferential reasoning.

The reconceptualization of pretests as learning events might even aid teachers in optimizing formative assessments, that is, informal assessments that are integrated into daily classroom practice. Although formative assessments are used widely to measure learning, sometimes at the beginning of a unit before instruction, they are less often directed toward directly improving student learning (Black & William, 1998). These embedded assessments are usually intended to allow teachers to better modulate their instruction to meet students' knowledge levels, but they might also serve as potent learning tools in their own right. One could imagine direct empirical tests of this speculation that would mirror the current experiments, but in a dynamic, interactive classroom setting.

### *Profiting From Standardized Testing*

Standardized testing is a more formal mode of assessment that plays an increasingly large role in classroom time. On the basis of the current No Child Left Behind Act, students are tested on their attainment of established curriculum standards for 2 weeks or more each year in many states, and that number is increasing as more districts seek to align with political pressures for assessment. Teachers and administrators alike describe these testing days as outside of instruction and as reducing an already affected curriculum schedule. Failed tests are viewed as indicating a lack of student progress and as a particularly egregious waste of needy students' time (e.g., Garrison, Jeung, & Inclán-Rodríquez, 2006; Hursh, 2007; Mathison & Freeman, 2003).

The current research lays the foundation for arguing that these testing days might be profitably integrated into the curriculum, and could actually facilitate subsequent learning for the unsuccessfully retrieved content. It is crucial, however, that after a test students be provided with an opportunity to restudy the tested material; a test that is not followed by instruction or feedback is likely to be of little use for items that were not answered correctly on the test. A recent study of standardized testing without aligned instruction suggests the same. When undergraduates and high school students were tested on retired SAT II questions without feedback, the tests led to an increase in posttest recall for participants who scored fairly well on the initial test, but led to no change in lower performing undergraduates and costs for high school students (Marsh, Agarwal, & Roediger, in press). Thus, in the absence of feedback or posttest learning opportunities, standardized testing may well be more problematic for lower performing students than higher performing students.

Reconsideration of these tests as learning events as well as assessments could have far-reaching implications for those reforms and interrelations with curriculum decisions. The alignment between tests and instruction has been the subject of substantial reform on the instructional side to ensure that instruction aligns with tested standards (e.g., see Amrein & Berliner, 2002; Herman & Golan, 1993; Stecher et al., 2008). Much less discussion has turned to tests' potential to directly affect students' learning and retention. Addressing this issue would require greater integration between testing and instruction, and such a shift could potentially address a common concern with high-stakes testing, namely, that the tests do not currently assess important aspects of the curriculum, thus reducing instructional depth.

Finally, another potential advantage of integrating standardized testing with instruction is self-regulatory and motivational, even in instances of failures. Although this was not the focus of the current analyses, "constructive failures" on test items may motivate learners and assist self-regulatory processes, such that learners become aware of what they do not currently know (e.g., Boekaerts, Pintrich, & Zeidner, 2000; McCaslin, 2006; Paris & Winograd, 1990; Paris, Byrnes, & Paris, 2001). Emphasizing the role of tests as learning events rather than as performance assessments may alleviate some of the pressure, and accompanying anxiety, that tests create for students when viewed solely as ultimate performance measures. In addition, reorienting to tests as learning tools could facilitate learners' ability to use them as prompts for deliberate practice as described by Ericsson, Krampe, and Tesch-Römer (1993).

### *Limitations and Future Directions*

The present studies are merely a first step in demonstrating the benefits of unsuccessful tests for future learning and, we hope, in encouraging educators and policymakers to consider the benefits of tests as learning events. In spite of the relatively clear results, several aspects of this study limit the breadth of interpretation for educational practice, and more must be done to establish the practical utility of this work. First, the current analyses exclusively focused on posttest scores for items that were answered incorrectly on a pretest. This allowed us to directly examine the impact of attempting yet failing to answer pretest questions, but may have underestimated the overall effects of testing. The extended study condition never had items removed, so these participants may have answered a small number of the posttest questions correctly regardless of their learning from the study materials. Future experiments should be conducted to develop a more precise measure of the effect size for providing pretest questions.

Second, future studies are necessary to ensure that the results generalize across diverse texts and are not tied to some particulars of the current experimental text. Third, we tested only fact-based questions; extending this research to additional types of questions would be useful in future studies. Fourth, the experimental materials were not embedded into participants' educational curricula more broadly, so there may be various differences in the way

participants engaged with the content, although we would anticipate that the cognitive underpinnings of testing and retrieval would remain constant.

Finally, before long-term gains can be made through changes to school practices, bridging studies must be conducted to better understand the relevant teacher-, school-, and district-level engagement with standardized and unit-level tests. The challenge is not only dissemination, but also of determining (a) how to effectively reframe testing as learning during everyday educational practice, and (b) whether modifying the well-established orientation to testing as a performance measure would lead to student gains. In short, the current results suggest that testing may have broad potential to directly enhance learning, whether or not the tests are successful, but, as with any finding in cognitive psychology, further research is necessary to demonstrate that such testing will effectively translate into classroom settings.

## Conclusion

When a learner makes an unsuccessful attempt to answer a question, both learners and educators often view the test as a failure, and assume that poor test performance is a signal that learning is not progressing. Thus, compared with presenting information to students, which is not associated with poor performance, tests can seem counterproductive. Tests are rarely thought of as learning events (Kornell & Bjork, 2007), and most educators would probably assume that giving students a test on material before they had learned it would have little impact on student learning beyond providing teachers with insight into their students' knowledge base. In terms of long-term learning, however, unsuccessful tests fall into the same category as a number of other effective learning phenomena (e.g., the spacing effect; see Dempster, 1988): Providing challenges for learners leads to low initial test performance, thereby alienating learners and educators, while simultaneously enhancing long-term learning (Bjork, 1994; Schmidt & Bjork, 1992). The current research suggests that tests can be valuable learning events, even if learners cannot answer test questions correctly, as long as the tested material has educational value and is followed by instruction that provides answers to the tested questions.

## References


Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18). Retrieved from http://epaa.asu.edu/epaa/v10n18/

Anderson, R. C., & Biddle, W. B. (1975). On asking people questions about what they are reading. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 9, pp. 89–132). New York: Academic Press.

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Bjork, R. A. (1988). Retrieval practice and the maintenance of knowledge. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory II* (pp. 396–401). London: Wiley.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80,* 139–144.

Boekaerts, M., Pintrich, P. R., & Zeidner, M. (2000). *Handbook of self-regulation.* San Diego, CA: Academic Press.

Bransford, J. S., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11,* 717–726.

Bransford, J. S., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education, 24,* 61–100.

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13,* 273–281.

Butler, A. C., & Roediger, H. L., III. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19,* 514–527.

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology, 19,* 619–636.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20,* 633–642.

Craig, S., Gholson, B., Ventura, M., Graesser, A. C., & the Tutoring Research Group. (2000). Overhearing dialogues and monologues in a virtual tutoring session: Effects on questioning and vicarious learning. *International Journal of Artificial Intelligence in Education, 11,* 242–253.

Craig, S. D., Sullins, J., Witherspoon, A., & Gholson, B. (2006). The deep-level-reasoning-question effect: The role of dialogue and deep-level-reasoning questions during vicarious learning. *Cognition and Instruction, 24,* 565–591.

Craik, F., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 67–84.

Davis, D., & Loftus, E. F. (2007). Internal and external sources of misinformation in adult witness memory. In M. P. Toglia, J. D. Read, D. F. Ross, & R. C. L. Lindsay (Eds.), *Handbook of eyewitness psychology (Vol. l). Memory for events* (pp. 195–237). Mahwah, NJ: Erlbaum.

Dempster, F. N. (1988). Informing classroom practice: What we know about several task characteristics and their effects on learning. *Contemporary Educational Psychology, 13,* 254–264.

Ericsson, K. A., Krampe, R. Th., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100,* 363–406.

Frase, L. T. (1968). Effect of question location, pacing, and mode upon retention of prose material. *Journal of Educational Psychology, 59,* 244–249.

Frase, L. T. (1970). Boundary conditions for mathemagenic behaviors. *Review of Educational Research, 40,* 337–347.

Garrison, C., Jeung, B., & Inclán-Rodríquez, R. (2006, March). Meeting English language learners' needs under No Child Left Behind. *Trends: Issues in Urban Education, 21,* 1–7.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 40,* 104.

Ghatala, E. S. (1981). The effect of internal generation of information on memory performance. *American Journal of Psychology, 94,* 443–450.

Ghatala, E. S. (1983). When does internal generation facilitate memory for sentences? *American Journal of Psychology, 96,* 75–83.

Gholson, B., & Craig, S. D. (2006). Promoting constructive activities that support vicarious learning during computer-based instruction. *Educational Psychology Review, 18,* 119–139.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81,* 392–399.

Glynn, S. M., Britton, B. K., & Tillman, M. H. (1985). Typographic cues in text: Management of the reader's attention. In D. Jonassen (Ed.), *The technology of text* (Vol. 2., pp. 192–209). Englewood Cliffs, NJ: Educational Technology Publications.

Glynn, S. M., & DiVesta, F. J. (1979). Control of prose processing via instructional and typographical cues. *Journal of Educational Psychology, 71,* 595–603.

Guthrie, E. R. (1952). *The psychology of learning* (rev. ed.). Oxford, England: Harper Bros.

Herman, J. L., & Golan, S. (1993). The effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice, 12,* 21–25, 41–42.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior, 10,* 562–567.

Huntley, J., & Davies, I. K. (1976). Preinstructional strategies: The role of pretests, behavioral objectives, overviews and advance organizers. *Review of Educational Research, 46,* 239–265.

Hursh, D. (2007). Exacerbating inequality: The failed promise of the No Child Left Behind Act. *Race Ethnicity and Education, 10,* 295–308.

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology, 83,* 340–344.

Kane, J. H., & Anderson, R. C. (1978). Depth of processing and interference effects in the learning and remembering of sentences. *Journal of Educational Psychology, 70,* 626–635.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19,* 528–558.

Karpicke, J. D., & Roediger, H. L., III. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57,* 151–162.

King, A. (1992). Comparison of self-questioning, summarizing, and notetaking-review as strategies for learning from lectures. *American Educational Research Journal, 29,* 303–323.

King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal, 31,* 338–368.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14,* 219–224.

Kornell, N., Hays, M. J., & Bjork, R. A. (in press). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology, 29,* 210–212.

Marsh, E. J., Agarwal, P. K., & Roediger, H. L. (in press). Memorial consequences of answering SAT II questions. *Journal of Educational Psychology: Applied.*

Marsh, E. J., Roediger, H. L., III, Bjork, R. A., & Bjork, E. L. (2007). Memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review, 14,* 194–199.

Mathison, S., & Freeman, M. (2003). Constraining elementary teachers' work: Dilemmas and paradoxes created by state mandated testing. *Education Policy Analysis Archives, 11*(34). Retrieved from http://epaa.asu.edu/epaa/images/logo.gif

Mayer, R. (1979). Twenty years of research on advance organizers: Assimilation theory is still the best predictor of results. *Instructional Science, 8,* 133–167.

Mayer, R. E. (2008). *Learning and instruction* (2nd ed.). Upper Saddle River, NJ: Merrill Prentice Hall.

McCaslin, M. (2006). Student motivational dynamics in the era of school reform. *The Elementary School Journal, 5,* 479–490.

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 371–385.

McDaniel, M. A., Roediger, H. L., III, & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14,* 200–206.

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors and feedback. *Psychonomic Bulletin & Review, 14,* 225–229.

No Child Left Behind Act of 2001, 20 U.S.C. §6319 (2008).

Paris, S. G., Byrnes, J. P., & Paris, A. H. (2001). Constructing theories, identities, and actions of self-regulated learners. In B. Zimmerman & D. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (2nd ed., pp. 253–287). Mahwah, NJ: Erlbaum.

Paris, S. G., & Winograd, P. (1990). How metacognition can promote learning and instruction. In B. F. Jones & L. Idol (Eds.), *Dimensions of thinking and cognitive instruction* (pp. 15–51). Hillsdale, NJ: Erlbaum.

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning* (Report No. NCER 2007–2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 3–8.

Pressley, M., Tanenbaum, R., McDaniel, M. A., & Wood, E. (1990). What happens when university students try to answer prequestions that accompany textbook material? *Contemporary Educational Psychology, 15,* 27–35.

Reynolds, R. E., & Anderson, R. C. (1982). Influence of questions on the allocation of attention during reading. *Journal of Educational Psychology, 74,* 623–632.

Reynolds, R. E., Standiford, S. N., & Anderson, R. C. (1979). Distribution of reading time when questions are asked about a restricted category of text information. *Journal of Educational Psychology, 71,* 183–190.

Richland, L. E., Bjork, R. A., & Linn, M. C. (2007). Cognition and instruction: Bridging laboratory and classroom settings. In F. Durso, R. Nickerson, S. Dumais, S. Lewandowsky, & T. Perfect (Eds.), *Handbook of applied cognition* (2nd ed., pp. 555–584). Chichester, England: Wiley.

Richland, L. E., Kao, L. S., & Kornell, N. (2008). Can unsuccessful tests enhance learning? In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society* (pp. 2338–2343). Mahwah, NJ: Erlbaum.

Rickards, J. P. (1976). Interaction of position and conceptual level of adjunct questions on immediate and delayed retention of text. *Journal of Educational Psychology, 68,* 210–217.

Rickards, J. P., & Hatcher, C. W. (1977–1978). Interspersed meaningful question learning questions as semantic cues for poor comprehenders. *Reading Research Quarterly, 13,* 538–553.

Roediger, H. L., III, & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181–210.

Roediger, H. L., III, & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17,* 249–255.

Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1155–1159.

Rothkopf, E. Z. (1965). Some theoretical and experimental approaches to problems in written instruction. In J. D. Krumboltz (Ed.), *Learning and the education process* (pp. 193–221). Chicago: Rand McNally.

Rothkopf, E. Z. (1966). Learning from written instructive materials: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal, 3,* 241–249.

Rothkopf, E. Z. (1982). Adjunct aids and the control of mathemagenic activities during purposeful reading. In W. Otto & S. White (Eds.), *Reading expository material.* New York: Academic Press.

Rothkopf, E. Z., & Bisbicos, E. E. (1967). Selective facilitative effects of interspersed questions on learning from written materials. *Journal of Educational Psychology, 58,* 56–61.

Sacks, O. (1995). *An anthropologist on Mars: Seven paradoxical tales by Oliver Sacks.* New York: Knopf.

Sagerman, N., & Mayer, R. E. (1987). Forward transfer of different reading strategies evoked by adjunct questions in science text. *Journal of Educational Psychology, 79,* 189–191.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3,* 207–217.

Skinner, B. F. (1958, October 24). Teaching machines. *Science, 128,* 969–977.

Snapp, J. C., & Glover, J. A. (1990). Advanced organizers and study questions. *Journal of Educational Research, 83,* 266–271.

Stecher, B. M., Epstein, S., Hamilton, L. S., Marsh, J. A., Robyn, A., McCombs, J. S., et al. (2008). *Pain and gain: Implementing No Child Left Behind in three states, 2004–2006.* Santa Monica, CA: Rand Corporation.

Terrace, H. S. (1963). Discrimination learning with and without "errors." *Journal of the Experimental Analysis of Behavior, 6,* 1–27.

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 6,* 175–184.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science, 31,* 1–60.

Waller, R. (1991). Typography and discourse. In R. Barr, D. R. Pearson, M. L. Kamil, & P. B. Mosenthal (Eds.), *Handbook of reading research* (Vol. 2, pp. 341–380). Mahwah, NJ: Erlbaum.

Watts, G. H., & Anderson, R. C. (1971). Effects of three types of inserted questions on learning from prose. *Journal of Educational Psychology, 62,* 387–394.

Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior, 16,* 465–478.

Yost, M., Avila, L., & Vexler, E. B. (1977). Effects of learning of post-instructional responses to questions of differing degrees of complexity. *Journal of Educational Psychology, 69,* 398–401.

## Appendix A

### Full List of Test Questions

1. What color is tomato juice to Mr. I?°^
2. How does Mr. I distinguish red and green traffic lights?*
3. It had been shown in the 1960s that there were cells in the primary visual cortex of monkeys (in the area termed $V_1$) that responded specifically to_____, but not to color.°
4. V4 specializes for responding to _____.
5. _____, in his famous prism experiment in 1666, showed that while light was composite—could be decomposed into, and recomposed by, all the colors of the spectrum.°
6. _____ _____, in 1802, feeling that there was no need to have an infinity of different receptors in the eye, each turned to a different wavelength postulated that 3 types of receptors would be enough.
7. What is total color blindness caused by brain damage called?°
8. Total color blindness caused by brain damage can reveal to us the mechanisms of _____ construction, specifically, here, how the brain "sees" (or makes) color.*
9. Color blindness, as ordinarily understood is something one is born with—a difficulty distinguishing red and green, or other colors, or (extremely rarely) an inability to see any colors at all, due to defects in color responding cells, the _____ of the retina.°
10. When given a large mass of yarns, containing 33 separate colors, how did he sort them?°

## Appendix B

### Replacement Items for Experiment 4

11. How does Mr. I distinguish flowers?°
12. Why was Mr. I stopped by the police when he decided to go to work again after the accident?°

## Appendix C

### Rewritten Items for Experiment 5

13. Tomato juice appears _____ to Mr. I.
14. There are cells in the primary visual cortex of monkeys that respond specifically to_____, but not to color.
15. Total color blindness caused by brain damage is called _____.
16. _____ showed that white light was composite.
17. Color blindness, as ordinarily understood, is something one is _____, rather than acquired later.
18. When given a large mass of yarns, containing 33 separate colors, Mr. I separated them by _____.
19. Mr. I distinguishes flowers by _____.
20. Mr. I was _____ when he decided to go to work again after the accident.

*Replaced for Experiments 3–5.

°Rewritten into standardized form for Experiment 5.

^Question numbering is for clarity and was not fixed in this order.

Received June 30, 2008
Revision received April 24, 2009
Accepted April 27, 2009 ■

*Research Article*

# Test-Enhanced Learning

## Taking Memory Tests Improves Long-Term Retention

**Henry L. Roediger, III, and Jeffrey D. Karpicke**

*Washington University in St. Louis*

**ABSTRACT**—*Taking a memory test not only assesses what one knows, but also enhances later retention, a phenomenon known as the testing effect. We studied this effect with educationally relevant materials and investigated whether testing facilitates learning only because tests offer an opportunity to restudy material. In two experiments, students studied prose passages and took one or three immediate free-recall tests, without feedback, or restudied the material the same number of times as the students who received tests. Students then took a final retention test 5 min, 2 days, or 1 week later. When the final test was given after 5 min, repeated studying improved recall relative to repeated testing. However, on the delayed tests, prior testing produced substantially greater retention than studying, even though repeated studying increased students' confidence in their ability to remember the material. Testing is a powerful means of improving learning, not just assessing it.*

In educational settings, tests are usually considered devices of assessment. Students take tests in class to assess what they have learned and take standardized tests like the SAT to assess their knowledge and aptitude. In many circumstances, such as university lecture courses, tests are given infrequently (often just two or three times a semester) and are generally perceived as a bother by faculty and students alike. We believe that the neglect of testing in all levels of education is misguided. To state an obvious point, if students know they will be tested regularly (say, once a week, or even every class period), they will study more and will space their studying throughout the semester rather than concentrating it just before exams (see Bangert-Drowns, Kulik, & Kulik, 1991; Leeming, 2002). However, more important for present purposes, testing has a powerful positive effect on future retention. If students are tested on material and successfully recall or recognize it, they will remember it better in the future than if they had not been tested. This phenomenon, called the testing effect, has been studied sporadically over a long period of time (e.g., Gates, 1917), but is not well known outside cognitive psychology.

Most experiments on the testing effect have been conducted in the verbal learning tradition using word lists (e.g., Hogan & Kintsch, 1971; Izawa, 1967; McDaniel & Masson, 1985; Thompson, Wenger, & Bartling, 1978; Tulving, 1967; Wheeler, Ewers, & Buonanno, 2003) or picture lists (Wheeler & Roediger, 1992) as materials. There have been a few experiments using materials found in educational contexts, beginning with Spitzer (1939; see too Glover, 1989, and McDaniel & Fisher, 1991). However, the title of Glover's article from 17 years ago still sums up the current state of affairs: "The 'testing' phenomenon: Not gone but nearly forgotten."

Our aim in the two experiments reported here was to investigate the testing effect under educationally relevant conditions, using prose materials and free-recall tests without feedback (somewhat akin to essay tests used in education). Most previous research has used tests involving recognition (like multiple-choice tests) or cued recall (like short-answer tests). A second purpose of our experiments was to determine whether testing facilitates learning beyond the benefits of restudying the material. In some testing-effect experiments, a study-test condition is compared with a study-only condition on a delayed retention test. When the subjects in the former condition outperform those in the latter on a final test, one can wonder whether the testing effect is simply due to study-test subjects being reexposed to the material during the test. It is no surprise that students will learn more with two presentations of material rather than one (although some of the word-list experiments cited earlier overcame this problem; see too Carrier & Pashler, 1992; Cull, 2000). To evaluate this restudying explanation of the testing effect, we had students in our control conditions restudy the entire set of material—which should, if anything, bias performance results in favor of this condition, because students who take free-recall tests (without feedback) can only reexperience whatever material they can recall.

Students in our experiments studied short prose passages covering general scientific topics. In Experiment 1, they either

Address correspondence to Henry L. Roediger, III, Department of Psychology, Washington University, Campus Box 1125, One Brookings Dr., St. Louis, MO 63130, e-mail: roediger@artsci.wustl.edu.

took a test on the material or studied it again before taking a final retention test 5 min, 2 days, or 1 week later. In Experiment 2, students studied a passage once and took three tests, studied three times and took one test, or studied the passage four times. They then took a final test 5 min or 1 week later. We predicted that performance on immediate retention tests would increase with the number of study opportunities, because massed practice typically produces short-term benefits (e.g., Balota, Duchek, & Paullin, 1989). However, we predicted that taking tests soon after studying would promote superior retention on delayed tests relative to repeatedly studying the material. This outcome would indicate that testing has positive effects on long-term retention above and beyond any effect of re-presentation of the material during the test.

## EXPERIMENT 1

### Method

*Subjects*

One hundred twenty Washington University undergraduates, ages 18 to 24, participated in partial fulfillment of course requirements.

*Materials*

Two prose passages were selected from the reading comprehension section of a test-preparation book for the Test of English as a Foreign Language (TOEFL; Rogers, 2001). Each passage covered a single topic ("The Sun" and "Sea Otters"), and each was divided into 30 idea units for scoring purposes. The passages were 256 and 275 words in length, respectively.

*Design*

A 2 × 3 mixed-factorial design was used. Learning condition (restudy vs. test) was manipulated within subjects, and delay of the final test (5 min, 2 days, or 1 week) was manipulated between subjects. The order of learning conditions (restudy or test) and the order of passages ("The Sun" or "Sea Otters") were counterbalanced across subjects.

*Procedure*

Subjects were tested during two sessions, in small groups (4 or fewer). They were told that Phase 1 consisted of four 7-min periods and that during any given period they would be asked to study one passage for the first time, restudy one of the passages, or take a recall test over one of the passages. During each study period, subjects read one passage for 7 min. During the test period, subjects were given a test sheet with the title of the to-be-recalled passage printed at the top and were asked to write down as much of the material from the passage as they could remember, without concern for exact wording or correct order. Subjects solved multiplication problems for 2 min between periods and for 5 min after the final period in Phase 1.

Phase 2 occurred after a 5-min, 2-day, or 1-week retention interval. In Phase 2, subjects were asked to recall the passages that they had learned in Phase 1. The recall instructions were identical to those given in Phase 1. Each retention test lasted 10 min, and subjects were instructed by the experimenter to draw a line on their test sheets to mark their place after each 1-min interval during the recall periods (Roediger & Thorpe, 1978). At the end of the experiment, subjects were debriefed and thanked for their participation.

### Results and Discussion

*Scoring*

Subjects' recall responses were scored by giving 1 point for each correctly recalled idea unit (out of 30). Initially, 40 recall tests were scored by two raters, and the Pearson product-moment correlation ($r$) between their scores was .95. Given the high interrater reliability, the remaining recall tests were scored by one rater.

*Initial Test*

On the initial 7-min test, subjects recalled on average 20.9 idea units, or approximately 70% of the passage. No differences were observed for the two passages or for the different counterbalancing orders.

*Final Test*

The mean proportion of idea units recalled on the final tests after the three retention intervals is shown in Figure 1. The cumulative recall data showed that subjects had exhausted their knowledge by the end of the retention interval and are not reported here. After 5 min, subjects who had studied the passage

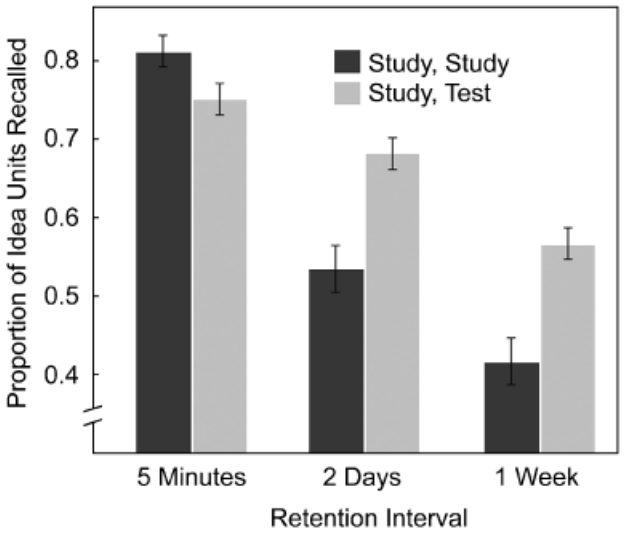


**Fig. 1.** Mean proportion of idea units recalled on the final test after a 5-min, 2-day, or 1-week retention interval as a function of learning condition (additional studying vs. initial testing) in Experiment 1. Error bars represent standard errors of the means.

twice recalled more than subjects who had studied once and taken a recall test. However, this pattern of results was reversed on the delayed tests 2 days and 1 week later. On these tests of long-term retention, subjects who had taken an initial test recalled more than subjects who had only studied the passages.

The results were submitted to a 2 × 3 analysis of variance (ANOVA), with learning condition (restudying or testing) and retention interval (5 min, 2 days, or 1 week) as independent variables. This analysis revealed a main effect of testing versus restudying, $F(1, 117) = 36.39$, $\eta_p^2 = .24$, which indicated that, overall, initial testing produced better final recall than additional studying. Also, the analysis revealed a main effect of retention interval, $F(2, 117) = 50.34$, $\eta_p^2 = .46$, which indicated that forgetting occurred as the retention interval grew longer. However, these main effects were qualified by a significant Learning Condition × Retention Interval interaction, $F(2, 117) = 32.10$, $\eta_p^2 = .35$, indicating that restudying produced better performance on the 5-min test, but testing produced better performance on the 2-day and 1-week tests.

Post hoc analyses confirmed that on the 5-min retention tests, restudying produced better recall than testing (81% vs. 75%), $t(39) = 3.22$, $d = 0.52$. However, the opposite pattern of results was observed on the delayed retention tests. After 2 days, the initially tested group recalled more than the additional-study group (68% vs. 54%), $t(39) = 6.97$, $d = 0.95$. The benefits of initial testing were also observed after 1 week: The tested group recalled 56% of the material, whereas the restudy group recalled only 42%, $t(39) = 6.41$, $d = 0.83$. Figure 1 depicts another interesting finding: The initially tested group recalled as much on the 1-week retention test as the additional-study group did after only 2 days (the initially tested group actually recalled slightly more). This surprising result indicates that taking an initial recall test prevented forgetting of information for an additional 5 days relative to repeated study.

Experiment 1 demonstrated that after an initial study episode, additional studying or testing had different effects on immediate and delayed final tests: Relative to testing, additional studying aided performance on immediate retention tests; in contrast, prior testing improved performance on delayed tests. The crossover interaction observed in Figure 1 is all the more impressive considering that no feedback was given on the tests. The testing effect on delayed retention tests is not simply due to reexposure to studied material during tests, but rather is due to some other process that has positive effects on retention. We consider candidate processes in the General Discussion.

## EXPERIMENT 2

In Experiment 2, we investigated the effects of repeated studying and repeated testing on retention, in part to replicate and extend the results of Experiment 1, but more to ask about effects of repeated testing. We were interested in the effects of repeated testing because most testing-effect experiments compare performance on final tests after subjects have or have not taken a single test earlier, as we did in Experiment 1. However, Wheeler and Roediger (1992) showed that taking three tests immediately after studying a list of pictures greatly improved retention on a final test relative to taking a single test or no test. In Experiment 2, we compared three conditions: Subjects studied a passage four times (and took no tests), studied it three times and took one test, or studied it once and took three tests. They then took a final test either 5 min or 1 week later. In addition to examining effects of repeated versus single tests, we made a few procedural changes to obtain estimates of how many times students in the various conditions actually read each passage. We also included a brief questionnaire after the initial learning session, asking subjects to rate how interesting and readable they found the passage and, more important, how well they thought they would remember it on a test 1 week later. We were particularly interested in subjects' predictions of how well they would remember the passage, because such judgments are not always correlated with actual performance (see Bjork, 1994; Koriat, Bjork, Sheffer, & Bar, 2004).

### Method

*Subjects and Materials*

One hundred eighty Washington University undergraduates, ages 18 to 24, participated in partial fulfillment of course requirements. The passages used in Experiment 1 were used again and were counterbalanced across conditions.

*Design*

A 3 × 2 between-subjects design was used. Subjects learned one of the two prose passages under one of three conditions (S = study, T = test): repeated study (SSSS), single test (SSST), or repeated test (STTT). Ninety subjects were given a final recall test following a 5-min retention interval, and 90 took a final test after 1 week. Thirty subjects were assigned to each of the six between-subjects conditions.

*Procedure*

The procedure used in Experiment 2 was similar to that used in Experiment 1. Subjects were again tested during two sessions, in small groups (4 or fewer). In Phase 1, they were told that they would be learning one passage during four consecutive periods. Subjects in the SSSS condition read the passage during four 5-min study periods; subjects in the SSST condition studied the passage during three periods and then took one recall test; those in the STTT condition studied the passage during one period and then took three consecutive recall tests. Students in the multiple-test condition were instructed to try hard to recall the entire passage on each successive test.

During study periods, subjects had 5 min to study the passage, and they recorded the number of times they read the entire passage by making tally marks on a separate sheet. During test

periods, subjects were given a blank sheet and were asked to recall as much of the material from the passage as they could remember, without concern for exact wording or correct order. Each test lasted 10 min, and subjects were instructed to draw a line on their test sheets to mark their place after each 1-min interval. Subjects solved multiplication problems for 2 min between periods and for 5 min after the final period in Phase 1.

At the end of Phase 1, subjects were given a questionnaire asking them to answer three questions using a 7-point scale. They indicated how interesting they thought the passage was (1 = *very boring*, 7 = *very interesting*), how readable they thought it was (1 = *very easy to read*, 7 = *very difficult to read*), and how well they thought they would remember the passage in 1 week (1 = *not very well*, 7 = *very well*). After completing the questionnaire, subjects in the 5-min retention-interval condition took the final recall test, and subjects in the 1-week condition were excused, returning for the final test 1 week later. The final recall test (Phase 2) was identical to the initial recall tests.

## Results and Discussion

### *Readings of the Passage*

The mean number of times subjects were able to read through the passage during each study period is presented in Table 1. No differences in these reading scores were observed for the two passages or for the 5-min and 1-week retention-interval groups. Across all conditions, subjects were able to read the entire passage approximately 3.5 times during a 5-min study period. The number of times subjects in the SSSS and SSST conditions read the passage increased slightly across consecutive study periods, $F(3, 177) = 1.62$, $\eta_p^2 = .03$, and $F(2, 118) = 4.99$, $\eta_p^2 = .08$, respectively. The reading scores in Table 1 simply illustrate that subjects read the passage many more times in the SSSS ($M = 14.2$) and SSST ($M = 10.3$) conditions than in the STTT ($M = 3.4$) condition.

### *Initial Tests*

Subjects in the STTT condition recalled 20.9, 21.2, and 21.1 idea units on each of the three initial recall tests, respectively, or about 70% of the passage in each case. No differences on the initial tests were observed for the two passages or for the 5-min and 1-week retention-interval groups. Measures of cumulative recall indicated that asymptotic levels of recall had been reached by the end of each test period. Subjects in the SSST condition recalled 23.1 idea units (77% of the passage) on their initial recall test. This was reliably greater recall than on the third test in the STTT condition, $t(118) = 3.17$, $d = 0.58$.

**TABLE 1**

***Mean Number of Times Subjects Were Able to Read the Entire Passage During the 5-Min Study Periods in Experiment 2***

| | Study period | | | | |
|---|---|---|---|---|---|
| Condition | 1 | 2 | 3 | 4 | Sum |
| SSSS | 3.4 | 3.5 | 3.6 | 3.7 | 14.2 |
| SSST | 3.2 | 3.5 | 3.6 | — | 10.3 |
| STTT | 3.4 | — | — | — | 3.4 |

**Note.** Condition labels indicate the order of study (S) and test (T) periods.

**TABLE 2**

***Mean Ratings on the Questionnaire Given After the Initial Learning Session in Experiment 2***

| | Rating | | |
|---|---|---|---|
| Condition | Interesting | Readable | Remember |
| SSSS | 3.8 | 2.5 | 4.8 |
| SSST | 4.1 | 2.5 | 4.2 |
| STTT | 4.6 | 2.8 | 4.0 |

**Note.** Condition labels indicate the order of study (S) and test (T) periods. Subjects rated how interesting the passage was (1 = *very boring*, 7 = *very interesting*), how readable the passage was (1 = *very easy to read*, 7 = *very difficult to read*), and how well they believed they would remember the passage in 1 week (1 = *not very well*, 7 = *very well*).

### *Questionnaire*

The mean ratings on the questionnaire given at the end of Phase 1 are displayed in Table 2. No differences in the questionnaire ratings were observed for the two passages or for the 5-min and 1-week retention-interval groups. Subjects in the SSSS condition rated the passage as less interesting than subjects in the SSST or STTT condition, $F(2, 177) = 3.88$, $\eta^2 = .04$, perhaps because of increased boredom with repeated readings. More interestingly, subjects in the SSSS condition were more confident that they would remember the passage in 1 week than were subjects in the SSST or STTT condition, $F(2, 177) = 6.09$, $\eta^2 = .06$. Post hoc analyses revealed that subjects in the SSSS condition predicted that they would remember the passage better than subjects in the SSST condition, $t(118) = 2.95$, $d = 0.54$, and subjects in the STTT condition, $t(118) = 3.35$, $d = 0.61$, but the latter two groups did not differ significantly in their predictions. The three groups did not differ in how they rated the readability of the passages ($F < 1$).

### *Final Tests*

The critical data are the mean proportions of idea units recalled on the final tests 5 min or 1 week later, displayed in Figure 2. The pattern of final test scores replicates the pattern of results found in Experiment 1. On the 5-min test, recall was correlated with repeated studying: The SSSS group recalled more than the SSST group (83% vs. 78%), who in turn recalled more than the STTT group (71%). However, on the 1-week test, recall was correlated with the number of tests given earlier: The STTT group recalled more than the SSST group (61% vs. 56%), who in turn recalled more than the SSSS group (40%).

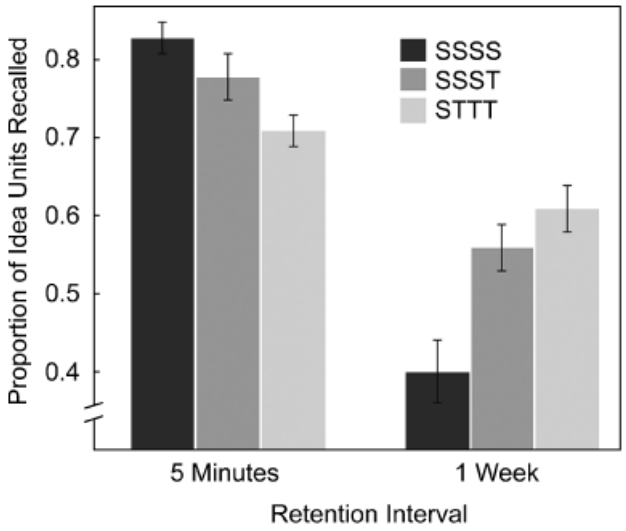


**Fig. 2.** Mean proportion of idea units recalled on the final test after a 5-min or 1-week retention interval as a function of learning condition (SSSS, SSST, or STTT) in Experiment 2. The labels for the learning conditions indicate the order of study (S) and test (T) periods. Error bars represent standard errors of the means.

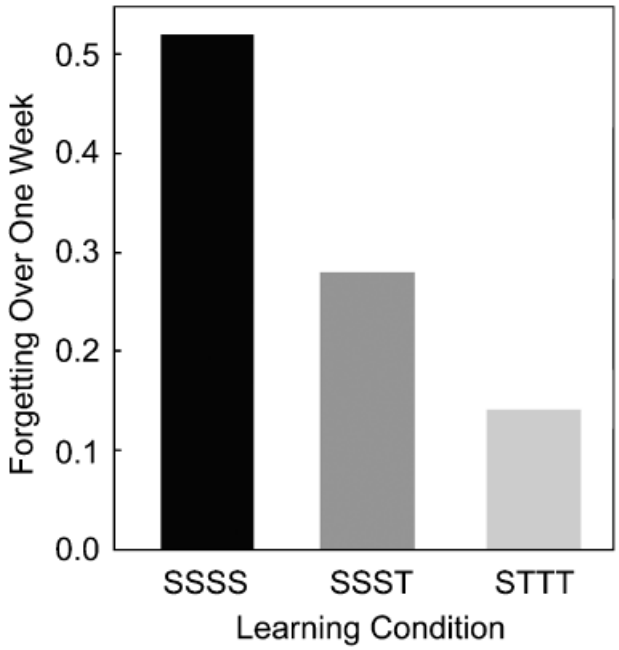


**Fig. 3.** Forgetting over 1 week as a function of learning condition (SSSS, SSST, or STTT) in Experiment 2. The labels for the learning conditions indicate the order of study (S) and test (T) periods.

The results in Figure 2 were submitted to a 2 × 3 ANOVA, with retention interval (5 min or 1 week) and learning condition (SSSS, SSST, or STTT) as independent variables. This analysis revealed a main effect of retention interval, $F(1, 174) = 122.53$, $\eta_p^2 = .41$, indicating that forgetting occurred. The effect of learning condition was only marginally significant, $F(2, 174) = 2.32$, $\eta_p^2 = .03$. However, these effects were qualified by a significant Learning Condition × Retention Interval interaction, $F(2, 174) = 18.48$, $\eta_p^2 = .18$. This interaction indicates that repeated studying produced short-term benefits, whereas repeated testing produced greater benefits on the delayed test.

Post hoc analyses confirmed these observations. On the 5-min retention test, subjects in the SSSS condition recalled more than subjects in the STTT condition, $t(58) = 4.70$, $d = 1.22$, as did subjects in the SSST condition, $t(58) = 2.24$, $d = 0.59$. However, this pattern was reversed on the 1-week retention test. Subjects in the STTT condition recalled more than subjects in the SSST condition, though this difference was marginal, $t(58) = 1.21$, $d = 0.31$. Subjects in the STTT condition also recalled significantly more than subjects in the SSSS condition, $t(58) = 4.78$, $d = 1.26$, as did subjects in the SSST condition, $t(58) = 3.21$, $d = 0.82$.

These data, involving absolute measures of forgetting, show greater forgetting in the pure-study condition than in the testing conditions. An alternative approach to examining forgetting is to use a proportional measure: (initial recall − final recall)/initial recall. In some studies, this alternative has led to different conclusions about rates of forgetting (Loftus, 1985). Proportional measures of forgetting are presented in Figure 3, in which it is obvious that subjects in the SSSS condition forgot far more (52%) than subjects in the SSST condition (28%) and than subjects in the STTT condition (14%). Subjects in the SSST condition forgot more than subjects in the STTT condition (28% vs. 14%). Thus, the proportional-forgetting analyses confirm those using the raw data and clearly demonstrate the powerful effect of repeated testing in preventing forgetting (cf. Wheeler & Roediger, 1992).

## GENERAL DISCUSSION

Both experiments showed the same pattern: Immediate testing after reading a prose passage promoted better long-term retention than repeatedly studying the passage. This outcome occurred even though the tests included no feedback. Clearly, the testing effect is not simply a result of students gaining reexposure to the material during testing, because restudying allowed students to reexperience 100% of the material but produced poor long-term retention (see too Wheeler et al., 2003). The positive effects of testing were dramatic: In Experiment 2, students in the repeated-testing condition recalled much more after a week than did students in the repeated-study condition (61% vs. 40%), even though students in the former condition read the passage only 3.4 times and those in the latter condition read it 14.2 times. Testing has a powerful effect on long-term retention.

The situation was different for tests taken shortly after learning: Repeated studying improved performance relative to repeated testing on final tests given after a 5-min retention interval, but the effect reversed on delayed tests. This pattern of results is analogous to the finding in the spacing-effect literature that massed presentation improves performance on immediate

tests, whereas spaced presentation leads to better performance on delayed tests (Balota et al., 1989; Peterson, Wampler, Kirkpatrick, & Saltzman, 1963). That is, in both cases, massed study leads to a short-term benefit, but the other manipulation (testing or spaced studying) has a greater effect on long-term retention. Both outcomes may reflect the role of desirable difficulties in promoting long-term retention (Bjork, 1994), as discussed later. This outcome on the immediate tests in the present experiments reveals just how powerful the testing effect is: Despite the benefits of repeated study shortly after learning, repeated testing produces strong positive effects on a delayed test.

Several overlapping theoretical approaches are useful in understanding our results. The findings are consistent with theories of transfer-appropriate processing that emphasize the compatibility between the operations engaged in during learning and testing phases (Morris, Bransford, & Franks, 1977; Roediger, 1990). The ability to remember a prose passage on a free-recall test a week after learning it is enhanced by practicing exactly this skill during learning. Practicing the skills during learning that are needed during retrieval generally enhances retention on both explicit and implicit memory tests (Roediger, Gallo, & Geraci, 2002). Although restudying the passages exposed students to the entire set of information, testing permitted practice of the skill required on future tests and hence enhanced performance after a delay.

McDaniel and his colleagues (McDaniel & Fisher, 1991; McDaniel, Kowitz, & Dunay, 1989; McDaniel & Masson, 1985) have argued that testing enhances learning by producing elaboration of existing memory traces and their cue-target relationships, and Bjork (1975, 1988) has suggested that testing operates by multiplying the number of "retrieval routes" to stored events. Bjork (1994, 1999) has also emphasized the need to introduce desirable difficulties into training and educational settings. Many study conditions and strategies that produce rapid learning and short-term benefits lead to poor long-term performance. Our results show that testing versus studying is another case in point: Testing clearly introduced a desirable difficulty during learning.

Relative to testing, repeated studying inflated students' confidence in their ability to remember the passages in the future, even though repeated-study subjects actually showed much poorer retention on delayed tests. Repeated studying is a strategy that students frequently report using and is often recommended to students by teachers (see Rawson & Kintsch, 2005, for discussion). Students may prefer repeated studying because it produces short-term benefits, and students often use ineffective learning strategies because they base their predictions of future performance on what produces rapid short-term gains. Although students in the repeated-study condition predicted they would perform very well a week later (relative to those in the other conditions), they actually performed the worst.

Free-recall testing even without feedback had large positive effects on retention in our experiments. Pashler and his colleagues (Pashler, Cepeda, Wixted, & Rohrer, 2005; Pashler, Zarow, & Triplett, 2003) have examined testing effects with feedback in paired-associates paradigms and also reported positive effects. Testing with feedback may improve performance even beyond the levels observed in the current research (McDermott, Kang, & Roediger, 2005). Judicious use of testing may improve performance in educational settings at all levels from elementary through university education, at least in fact-based courses (Roediger & Karpicke, 2006). Frequent testing leads students to space their study efforts, permits them and their instructors to assess their knowledge on an ongoing basis, and—most important for present purposes—serves as a powerful mnemonic aid for future retention. The boundary conditions for the testing effect are not yet known, but we suspect that tests will produce strong effects when they occur relatively soon after learning and permit relatively high levels of performance (Balota, Duchek, Sergent-Marshall, & Roediger, in press; Landauer & Bjork, 1978; Logan & Balota, 2005; Spitzer, 1939). We believe the time is ripe for a thorough examination of the mnemonic benefits of testing and its potentially important consequences for improving educational practice.

***Acknowledgments***—This research was supported by a grant from the Institute of Educational Sciences and by a Collaborative Activity Grant from the James S. McDonnell Foundation. We thank Jane McConnell for her help and John Wixted for comments.

## REFERENCES

Balota, D.A., Duchek, J.M., & Paullin, R. (1989). Age-related differences in the impact of spacing, lag, and retention interval. *Psychology and Aging*, *4*, 3–9.

Balota, D.A., Duchek, J.M., Sergent-Marshall, S., & Roediger, H.L., III. (in press). Does expanded retrieval produce benefits over equal interval spacing? Explorations in healthy aging and early stage Alzheimer's disease. *Psychology and Aging*.

Bangert-Drowns, R.L., Kulik, J.A., & Kulik, C.L.C. (1991). Effects of frequent classroom testing. *Journal of Educational Research*, *85*, 89–99.

Bjork, R.A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R.L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Bjork, R.A. (1988). Retrieval practice and the maintenance of knowledge. In M.M. Gruneberg, P.E. Morris, & R.N. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (Vol. 1, pp. 396–401). New York: Wiley.

Bjork, R.A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Bjork, R.A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*, 633–642.

Cull, W.L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, *14*, 215–235.

Gates, A.I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, *6*(40).

Glover, J.A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*, 392–399.

Hogan, R.M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, *10*, 562–567.

Izawa, C. (1967). Function of test trials in paired-associate learning. *Journal of Experimental Psychology*, *75*, 194–209.

Koriat, A., Bjork, R.A., Sheffer, L., & Bar, S.K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General*, *133*, 643–656.

Landauer, T.K., & Bjork, R.A. (1978). Optimum rehearsal patterns and name learning. In M.M. Gruneberg, P.E. Morris, & R.N. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). London: Academic Press.

Leeming, F.C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology*, *29*, 210–212.

Loftus, G. (1985). Evaluating forgetting curves. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 397–406.

Logan, J.M., & Balota, D.A. (2005). *Spaced and expanded retrieval effects in younger and older adults*. Manuscript submitted for publication.

McDaniel, M.A., & Fisher, R.P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology*, *16*, 192–201.

McDaniel, M.A., Kowitz, M.D., & Dunay, P.K. (1989). Altering memory through recall: The effects of cue-guided retrieval processing. *Memory & Cognition*, *17*, 423–434.

McDaniel, M.A., & Masson, M.E.J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 371–385.

McDermott, K.B., Kang, S., & Roediger, H.L., III. (2005, January). *Test format and its modulation of the testing effect*. Paper presented at the biennial meeting of the Society for Applied Research in Memory and Cognition, Wellington, New Zealand.

Morris, C.D., Bransford, J.D., & Franks, J.J. (1977). Levels of processing versus transfer-appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519–533.

Pashler, H., Cepeda, N.J., Wixted, J.T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 3–8.

Pashler, H., Zarow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1051–1057.

Peterson, L.R., Wampler, R., Kirkpatrick, M., & Saltzman, D. (1963). Effect of spacing of presentations on retention of paired-associates over short intervals. *Journal of Experimental Psychology*, *66*, 206–209.

Rawson, K.A., & Kintsch, W. (2005). Rereading effects depend on time of test. *Journal of Educational Psychology*, *97*, 70–80.

Roediger, H.L., III. (1990). Implicit memory: Retention without remembering. *American Psychologist*, *45*, 1043–1056.

Roediger, H.L., III, Gallo, D.A., & Geraci, L. (2002). Processing approaches to cognition: The impetus from the levels-of-processing framework. *Memory*, *10*, 319–332.

Roediger, H.L., III, & Karpicke, J.D. (2006). *The power of testing memory: Implications for educational practice*. Unpublished manuscript, Washington University in St. Louis.

Roediger, H.L., III, & Thorpe, L.A. (1978). The role of recall time in producing hypermnesia. *Memory & Cognition*, *6*, 296–305.

Rogers, B. (2001). *TOEFL CBT Success*. Princeton, NJ: Peterson's.

Spitzer, H.F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*, 641–656.

Thompson, C.P., Wenger, S.K., & Bartling, C.A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 210–221.

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *6*, 175–184.

Wheeler, M.A., Ewers, M., & Buonanno, J. (2003). Different rates of forgetting following study versus test trials. *Memory*, *11*, 571–580.

Wheeler, M.A., & Roediger, H.L., III. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science*, *3*, 240–245.

(Received 2/4/05; Accepted 3/24/05;
Final materials received 4/13/05)

# β-Amyloid accumulation in the human brain after one night of sleep deprivation

Ehsan Shokri-Kojori[a,1], Gene-Jack Wang[a,1], Corinde E. Wiers[a], Sukru B. Demiral[a], Min Guo[a], Sung Won Kim[a], Elsa Lindgren[a], Veronica Ramirez[a], Amna Zehra[a], Clara Freeman[a], Gregg Miller[a], Peter Manza[a], Tansha Srivastava[a], Susan De Santi[b], Dardo Tomasi[a], Helene Benveniste[c], and Nora D. Volkow[a,1]

[a]Laboratory of Neuroimaging, National Institute on Alcohol Abuse and Alcoholism, National Institutes of Health, Bethesda, MD 20892; [b]Piramal Pharma Inc., Boston, MA 02108; and [c]Department of Anesthesiology, Yale School of Medicine, New Haven, CT 06510

Edited by Michael E. Phelps, University of California, Los Angeles, CA, and approved March 13, 2018 (received for review December 14, 2017)

**The effects of acute sleep deprivation on β-amyloid (Aβ) clearance in the human brain have not been documented. Here we used PET and $^{18}$F-florbetaben to measure brain Aβ burden (ABB) in 20 healthy controls tested after a night of rested sleep (baseline) and after a night of sleep deprivation. We show that one night of sleep deprivation, relative to baseline, resulted in a significant increase in Aβ burden in the right hippocampus and thalamus. These increases were associated with mood worsening following sleep deprivation, but were not related to the genetic risk (APOE genotype) for Alzheimer's disease. Additionally, baseline ABB in a range of subcortical regions and the precuneus was inversely associated with reported night sleep hours. APOE genotyping was also linked to subcortical ABB, suggesting that different Alzheimer's disease risk factors might independently affect ABB in nearby brain regions. In summary, our findings show adverse effects of one-night sleep deprivation on brain ABB and expand on prior findings of higher Aβ accumulation with chronic less sleep.**

beta amyloid | sleep | hippocampus | Alzheimer's disease | glymphatic system

Beta-amyloid (Aβ) is present in the brain's interstitial fluid (ISF) and is considered a metabolic "waste product" (1). Mechanisms by which Aβ is cleared from the brain are not completely understood (2), although there is evidence that sleep plays an important role in Aβ clearance (3). In rodents, chronic sleep restriction led to increases in ISF Aβ levels (4) and in a *Drosophila* model of Alzheimer's disease (AD), chronic sleep deprivation (SD) resulted in higher Aβ accumulation (5). In healthy humans, imaging studies have revealed associations between self-reports of less sleep duration or poor sleep quality and higher Aβ burden (ABB) in the brain (6–8), which is a risk factor for AD. This association has been considered bidirectional because increased ABB could also lead to impairments in sleep (9, 10). Notably, increased ABB in the brain has been associated with impairment of brain function (11, 12). Thus, strategies that prevent Aβ accumulation in the brain could promote healthy brain aging and be useful in preventing AD. In this respect, there is increasing evidence that sleep disturbances might contribute to AD, in part by facilitating accumulation of Aβ in the brain (13).

To better characterize ABB dynamics, studies have focused on the effects of sleep patterns on ABB in the CNS. In rodents, it has been shown that Aβ clearance from the brain's ISF predominately occurred during sleep (4), which was ascribed to the glymphatic pathway, operating most efficiently during sleep (3, 14, 15). Clinical studies have also shown that Aβ levels in the cerebrospinal fluid (CSF) are the highest before sleep and the lowest after wakening, while CSF Aβ clearance was counteracted by SD (16). However, there are some inconsistencies between animal models and findings in humans (17), and Aβ increases in human CSF could reflect factors other than ABB increases in the brain itself (18–21). Notably, the effects of acute SD on Aβ clearance in the human brain have not been documented. This observation will be important for understanding the contribution of sleep to Aβ clearance from the brain and the regional specificity of such effects.

Here we evaluated the effects of one-night SD on ABB in healthy controls to investigate whether sleep affects clearance of Aβ from the human brain. For this purpose, we used positron emission tomography (PET) with which it is now possible to measure ABB in the living human brain. There are several validated PET radiotracers for this purpose, including $^{18}$F-florbetaben (FBB) (22, 23). It is believed that such radiotracers predominantly bind to insoluble $A\beta_{42}$ plaques (24–27), but there is recent evidence that they also bind to soluble $A\beta_{42}$ forms (28). Thus, we reasoned that PET and FBB could be used to detect increases in ABB because of acute SD, directly in the human brain (3). First, we aimed to assess the effect of one-night SD on brain ABB with PET-FBB in healthy controls ($n = 20$, 22–72 y old, 10 females) (Table S1), and compared the measures to baseline brain ABB captured at the same time of the day but following a night of rested sleep [referred to as rested-wakefulness (RW)]. Second, we aimed to replicate in our sample the previously reported association between sleep history and brain ABB (when measured after RW) (6–8). For our first aim, we hypothesized that one night of SD would increase ABB in the hippocampus, which shows some of the earliest structural and functional changes in AD (29, 30). For our second aim, we hypothesized that history of poor sleep would be associated with

## Significance

There has been an emerging interest in sleep and its association with β-amyloid burden as a risk factor for Alzheimer's disease. Despite the evidence that acute sleep deprivation elevates β-amyloid levels in mouse interstitial fluid and in human cerebrospinal fluid, not much is known about the impact of sleep deprivation on β-amyloid burden in the human brain. Using positron emission tomography, here we show that acute sleep deprivation impacts β-amyloid burden in brain regions that have been implicated in Alzheimer's disease. Our observations provide preliminary evidence for the negative effect of sleep deprivation on β-amyloid burden in the human brain.

Author contributions: N.D.V. conceived study; E.S.-K., G.-J.W., C.E.W., S.B.D., S.W.K., S.D.S., D.T., H.B., and N.D.V. designed research; E.S.-K., G.-J.W., C.E.W., S.B.D., M.G., S.W.K., E.L., V.R., A.Z., C.F., G.M., P.M., T.S., S.D.S., D.T., H.B., and N.D.V. performed research; E.S.-K. analyzed data; and E.S.-K. and N.D.V. wrote the paper.

Conflict of interest statement: S.D.S. was an employee of Piramal Pharma Inc., which partly supported the radiotracer for this study.

This article is a PNAS Direct Submission.



[1]To whom correspondence may be addressed. Email: ehsan.shokrikojori@nih.gov, genejack.wang@nih.gov, or nvolkow@nida.nih.gov.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1721694115/-/DCSupplemental.

Published online April 9, 2018.

higher ABB in the hippocampus, precuneus, and medial prefrontal cortex (6, 8, 31).

## Results

**Acute SD Effects.** To compare the differences in FBB binding [quantified as relative standard uptake value (SUVr) and used as a marker of ABB] (*Methods*) after acute SD versus that obtained after RW, we used a voxelwise paired $t$ test in statistical parametric mapping (SPM) (*Methods*). This analysis showed that images obtained after SD compared with those obtained after RW had significantly higher FBB binding (ABB increases) in a right lateralized cluster (Fig. 1*A*) that comprised hippocampal, parahippocampal, and thalamic regions (Table S2). Of note, the increases in FBB SUVr in this cluster were robust and observed in 19 of 20 participants (Fig. 1*B*) from RW (mean = 1.35, standard deviation = 0.06) to SD (mean = 1.42, standard deviation = 0.07; a 5% increase, $P < 0.0001$). To further confirm this finding, we quantified FBB SUVr in an a priori hippocampal region of interest (ROI) (*Methods* and Fig. 1*D*) and compared the measures after SD to those after RW. This ROI analysis also showed a significant increase in FBB SUVr in the right hippocampus ROI from RW to SD ($P = 0.046$, two-tailed, Cohen's $d = 0.48$) but not in the left hippocampus ($P = 0.4$). The magnitude of the Aβ changes in the hippocampal cluster varied significantly between subjects (Fig. 1*B*) (−0.58% to +16.1%). We found that this variability was not associated with gender, age, or apolipoprotein E (APOE)-based odds ratio for AD (ORAD) (*Methods*) ($P > 0.3$). Notably, changes in FBB SUVr in the subcortical cluster were significant in both males ($P = 0.0008$) and females ($P = 0.003$). In addition, reported sleep hours (SH) and total score (TS) for sleep quality (Pittsburgh Sleep Questionnaire Inventory, PSQI) (*Methods*) were not associated with these SD-related increases. Thus, the mechanisms accounting for the observed between-subject variability are still unclear. Subjective behavioral assessment revealed that SD negatively impacted mood compared with RW (*Methods* and Fig. S1). We assessed if the effects of SD on mood were correlated with increases in ABB in the right hippocampal cluster. This analysis showed that mood worsening was negatively associated with changes in FBB SUVr [$r$ (16) = −0.50, $P = 0.03$] (Fig. 1*C*) such that participants with larger increases in ABB in the hippocampal cluster (Fig. 1*A*) had more mood worsening after SD. Because the quantification of ABB using FBB SUVr can be sensitive to blood perfusion effects, we quantified FBB accumulation using measures of binding potential (BPnd), which are less sensitive to blood perfusion effects than SUVr measures. BPnd in the cluster where we observed the SD effect (*Methods* and Fig. 1*A*) was also significantly higher in SD relative to RW [$t(19) = 3.57$, $P = 0.002$] (Fig. S2*A*). Moreover, SD-related changes in BPnd were significantly correlated with those observed with FBB SUVr [$r(18) = 0.53$, $P = 0.016$] (Fig. S2*B*), further supporting that SD-related increases in FBB SUVr are not primarily driven by perfusion effects.

**Sleep, APOE, and ABB.** Prior studies had reported an association between reported SH and sleep quality and (cortical) ABB in healthy middle-aged and older individuals (6–8). We tested whether we would corroborate those observations in our sample using the measures obtained during RW. We found that reported average SH inversely correlated with FBB SUVr [$r(18) = -0.5$, $P = 0.024$] and with BPnd [$r(18) = -0.57$, $p = 0.009$] at RW in the subcortical cluster that showed increases in ABB with SD (Fig. 2*A*), thus supporting long-term susceptibility of these regions to increased ABB with less SH. Voxelwise regression analysis of FBB SUVr on SH showed that less SH was associated with higher FBB SUVr in the bilateral putamen, parahippocampus, and right precuneus (Fig. 2*B* and Table S3). Interestingly, the SH-related brain areas were more extensive than the areas associated with (acute) SD-induced ABB increases. For regional ABB, SH-related subcortical regions (Table S3) had minimal overlap with subcortical areas related to ORAD (Fig. 2*B*), which included the bilateral lentiform nucleus and pallidum (Table S4). These observations suggested that different brain

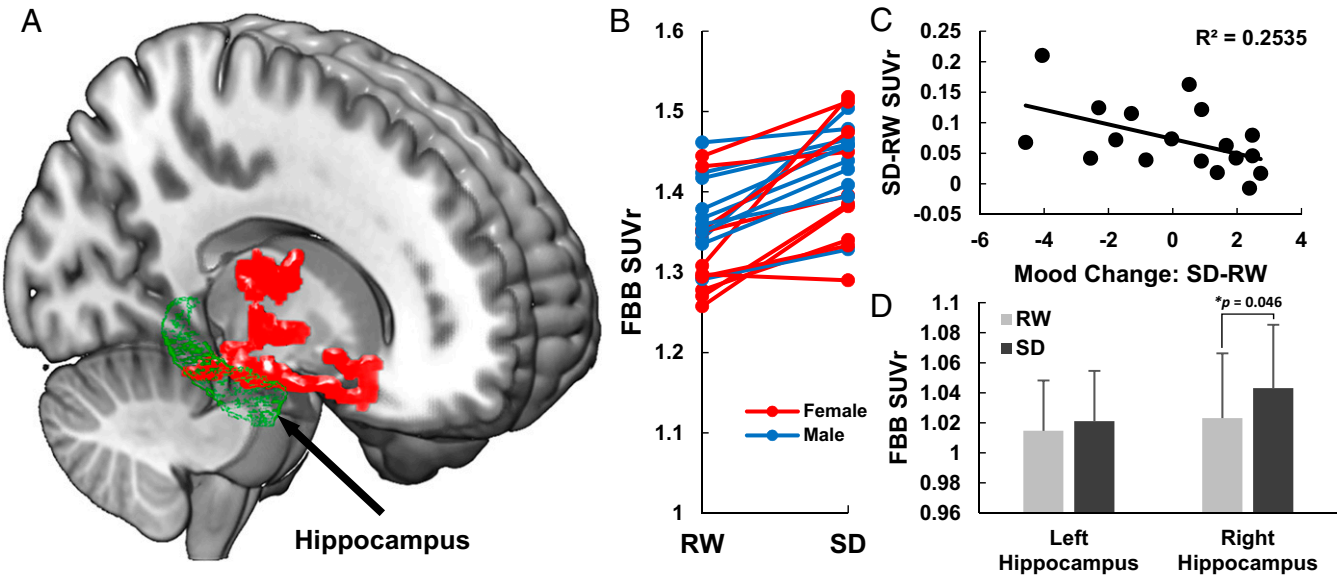


**Fig. 1.** Effects of one-night SD on ABB. (*A*) Voxelwise paired $t$ test between RW and SD conditions highlighting the hippocampus as well as other subcortical structures ($P_{FWE} < 0.05$, cluster-size corrected) (Table S1). (*B*) Subject-level changes in FBB SUVr (in the red cluster identified in *A*) from RW to SD. There was no significant effect of gender, or gender × sleep interaction ($P > 0.15$). (*C*) Association between changes in mood from RW to SD and changes in the FBB SUVr for the cluster identified in *A*. Mood change was quantified using the principal component of the changes in self-report measures from RW to SD, which accounted for 35.5% of the variance. Self-report measures of alert, friendly, happy, social, and energetic significantly decreased, and measures of tired and difficulty staying awake significantly increased from RW to SD ($P < 0.001$, two-tailed) (see also Fig. S1). (*D*) Average FBB SUVr in a priori hippocampus ROIs across subjects. Error bars show standard deviation (*Methods*).

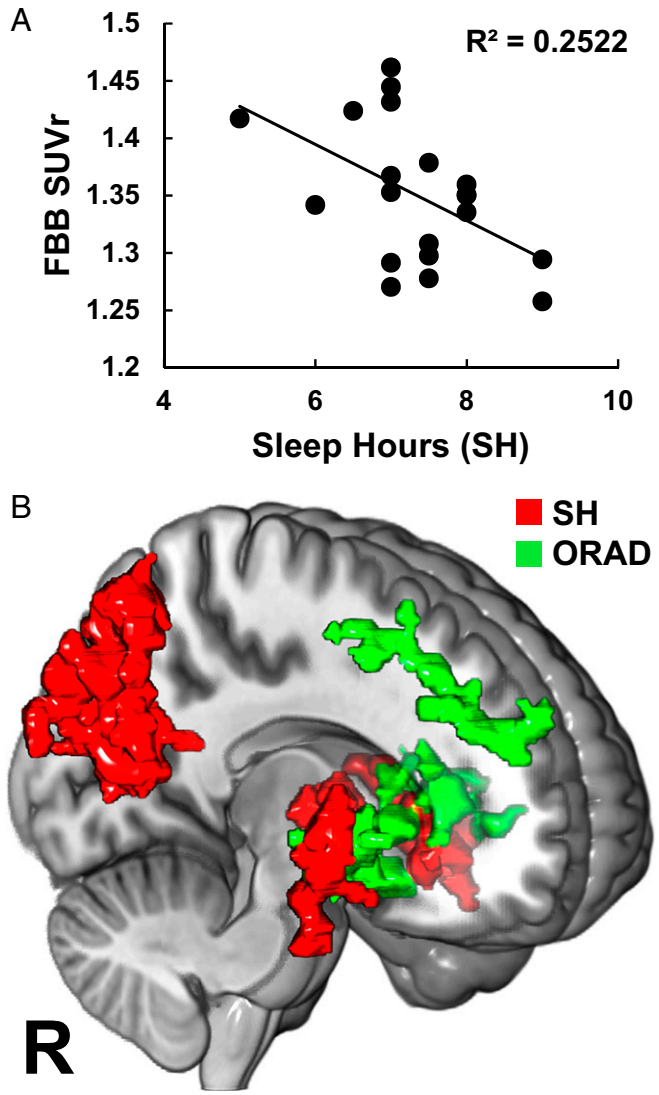


**Fig. 2.** Relationship between SH, APOE genotype, and ABB. (*A*) Regression of FBB SUVr (indexing ABB) for the red cluster shown in Fig. 1*A* at RW against SH. (*B*) Three-dimensional rendering of the areas showing an association between higher FBB SUVr at RW and lower SH (red clusters, $P_{FWE} < 0.05$, cluster-size corrected) (Table S3), as well as areas showing an association between higher FBB SUVr and higher APOE-based genetic risk for AD (quantified as log of ORAD) (green clusters, $P_{FWE} < 0.05$, cluster-size corrected, except the right-sided subcortical cluster, which was cluster size-corrected at $q_{FDR} < 0.05$) (Table S4). The subcortical clusters related to SH and ORAD had minimal spatial overlap (18 voxels, <3%) (Tables S3 and S4). Notably, FBB SUVr within the SH-related subcortical clusters (red, *B*) was not associated with ORAD ($P = 0.25$) and FBB SUVr within the ORAD-related subcortical clusters (green, *B*) was not associated with SH ($P = 0.2$).

regions could be independently affected by different AD-risk factors (i.e., sleep vs. APOE). This observation is consistent with prior findings reporting that the APOE genotype did not moderate the relationship between sleep measures and ABB in the brain (7).

## Discussion

Our findings provide preliminary evidence for the role of SD on Aβ accumulation in the human brain. The increases in ABB after SD in the hippocampus, which is considered among the most-sensitive brain regions to AD neuropathology (29), is consistent with epidemiological data identifying impaired sleep as a risk factor for AD (9, 10) and with recent evidence showing that disruption of deep sleep increases Aβ in human CSF (32). We also showed Aβ increases in the thalamus after SD, which is a brain region that shows increases in Aβ in the early stages of AD (33). The regional increases in ABB that we observed from RW to SD might reflect decreased clearance of Aβ, presumably from lack of sleep, thus supporting the role of glymphatic system in clearing Aβ from the brain during sleep. While this effect was observed in rodents (3), no study as of now has been able to directly measure Aβ clearance reflecting glymphatic function in the human brain. Other mechanisms have been shown to stimulate Aβ clearance during sleep, such as γ-oscillations during the rapid eye-movement cycle (34). Alternatively, Aβ increases in the hippocampus after SD could reflect increases in Aβ synthesis associated with endogenous neuronal activity during SD (35), or sleep-related changes in hippocampal neurogenesis (36, 37). While we are interpreting our findings of increases in ABB after SD to reflect Aβ accumulation due to lack of glymphatic clearance (or other unknown mechanisms), we cannot rule out the possibility that they reflect increases in the synthesis of Aβ (38) from lack of sleep. Preclinical studies that monitor Aβ clearance during sleep using PET radiotracers alongside optical imaging of ISF are needed to corroborate whether Aβ ligands and PET have potential as markers of glymphatic function in the human brain. Future work should also study the extent that elevated ABB in the brain is related to increases in Aβ levels in the peripheral tissue following SD (18, 20).

Detection of Aβ plaques with PET is clinically relevant for the diagnosis of AD, while elevated brain ABB in mild cognitive impairment (MCI) has been suggested as a risk factor for progression to AD (22, 39). ABB increases as a function of aging and AD severity, with an estimated 17% increase from younger to older adults (40). Relative to healthy elderly, estimated increases of 21% and 43% have been reported in individuals with MCI and AD, respectively (40). In our study, the increases in ABB due to one-night SD were smaller (5%) (Fig. 1 *A* and *B*) and were observed in a subcortical cluster that included the hippocampus. At this stage, we are uncertain whether such SD-related increases in ABB may subside following rested sleep. In addition, the magnitude of the observed effects should be interpreted with caution considering the methodological limitations and potential confounds, such as the effects of blood-flow changes and the limited sensitivity of current PET radiotracer to soluble Aβ (28). While, prior studies have reported associations with chronic poor sleep and higher average ABB across large a priori-selected cortical areas (6–8), our results highlight the relevance of studying subcortical regions for the associations between ABB and sleep.

It has been shown in rodents that Aβ plaques can form very rapidly (41), while Aβ plaques in their earliest stages show the highest levels of neuritic dystrophy (42), thus suggesting that increases in ABB due to one-night SD could adversely impact the brain. Chronic poor sleep could then result in higher baseline ABB levels, helping to explain the association with higher ABB at RW (in the cluster showing SD effects) (Fig. 1*A*) and less reported SH (Fig. 2*A*). In addition, SH negatively correlated with ABB in the precuneus, putamen, and parahippocampus (Fig. 2*B* and Table S3). These findings are consistent with studies showing that chronic sleep restriction lead to elevated cortical Aβ oligomers levels (that are neurotoxic) in the mouse cortex (43), which might in part reflect up-regulation of β-secretase 1 (44). However, a confounder is that increased ABB could in turn exacerbate sleep problems (10).

It is noteworthy that ABB increases in the hippocampus are less pronounced in MCI (45) or early-stage AD than in the precuneus (22). Instead, changes in hippocampal volume, glucose metabolism (30), and the blood–brain barrier (46) were most evident in patients with early-stage AD. Thus, it is possible

that the increases in ABB in the hippocampus with sleep disruptions might trigger local neurotoxicity without necessarily resulting in marked plaque accumulation. The precuneus was not significantly affected by SD; however, we found an association between reported SH and ABB in this region (Fig. 2*B* and Table S3), again suggesting that distinct processes might mediate the effects of acute vs. chronic SD on regional ABB. Future work should investigate the extent to which the effects of acute SD and less SH might reflect distinct mechanisms (i.e., neuroinflammation triggered by chronic poor sleep) (47, 48) or sensitivity of FBB to different forms of Aβ (soluble oligomers vs. plaques) (28). One limitation of this work includes the inability of PET-FBB to distinguish soluble from insoluble Aβ (27, 28, 49). We suggested that the interruption of glymphatic clearance by SD would increase ABB in the brain, yet our findings do not demonstrate the mechanisms that account for the Aβ accumulation with SD. In our study, estimates of ABB were obtained using FBB SUVr, which is sensitive to physiological factors such as blood flow, which could have been impacted by SD (50, 51). However, recent evidence suggests that blood-flow effects are small on FBB SUVr estimated with later time points (52) (*Methods*). We also corroborated the FBB SUVr findings with BPnd measures that are less sensitive to blood-flow effects (Fig. S2).

In our study, we did not predict the laterality effect for the SD-related increases in ABB, and while this could reflect the sensitivity of the glymphatic system to orientation of the head during sleep (14), we did not record head position. Consistent with the recognized role of the hippocampus and thalamus in mood disorders (53), the association between SD-related increases in ABB and mood worsening (Fig. 1*C*) supported the functional significance of elevated ABB. This association could reflect the previously reported contribution of the hippocampus in modulating mood changes that follow SD (54). The effects of SD on the hippocampus have also been implicated in the memory impairment associated with SD, although we did not measure effects of SD on memory in our study. Even though our sample was small ($n = 20$), we were able to identify a significant effect of SD on brain ABB with no significant interaction with gender. Because of the small sample size of our study, future studies are needed to assess the generalizability to a larger and more diverse population and to more reliably characterize potential gender effects.

In summary, this study documents an effect of one-night SD on ABB in the hippocampus, thus providing preliminary evidence that sleep, among other factors, could influence Aβ clearance in the human brain. Our results highlight the relevance of good sleep hygiene for proper brain function and as a potential target for prevention of AD (31, 55).

## Methods

**Participants.** Twenty-two healthy individuals were recruited at the National Institutes of Health, of which 20 (10 females, age: 39.8 y ± 10.4, range: 22–72 y old) completed two PET scan sessions to measure ABB. All participants provided informed consent to participate in the study that was approved by the Institutional Review Board at the NIH (Combined Neurosciences White Panel). Exclusion criteria were: (*i*) urine positive for psychotropic drugs; (*ii*) history of alcohol or drug use disorders; (*iii*) present or past history of neurological or psychiatric disorder, including evidence of cognitive impairment; (*iv*) use of psychoactive medications in the past month (i.e., opiate analgesics, stimulants, sedatives); (*v*) currently taking prescription medications (i.e., antihistamines, antihypertensive, antibiotics); (*vi*) medical conditions that may alter cerebral function; (*vii*) cardiovascular and metabolic diseases; and (*viii*) history of head trauma with loss of consciousness longer than 30 min. Table S1 summarizes physiological and neuropsychological assessment to ensure participants were healthy and were not cognitively impaired.

**Structural MRI.** Participants also underwent MRI in a 3.0 T Magnetom Prisma scanner (Siemens Medical Solutions) using a 32-channel head coil to collect T1-weighted 3D MPRAGE (TR/TE = 2,400/2.24 ms, 0.8-mm isotropic resolution) and T2-weighted spin-echo multislice (TR/TE = 3,200/564 ms, 0.8-mm in-plane resolution). MRI was processed using the minimal preprocessing pipeline of the Human Connectome Project (56). Specifically, FreeSurfer v5.3 (Martinos Center for Biomedical Imaging; https://surfer.nmr.mgh.harvard.edu/) was used for anatomical data segmentation. In addition, each MRI image underwent gradient distortion correction, field map processing, spatial normalization to the stereotactic space of the Montreal Neurological Institute (MNI) with 2-mm isotropic resolution, and brain masking using routines from University of Oxford's Center for Functional Magnetic Resonance Imaging of the Brain Software Library release 5.0 (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki). FreeSurfer segmentation ('wmparc.nii') in the MNI space was used for generating a subject-specific mask of the hippocampus (label numbers: 17 and 53 for left and right hippocampus, respectively).

**PET Data Acquisition.** The PET scans were performed using a high-resolution research tomography Siemens scanner on two separate scan days with FBB. FBB was injected through an intravenous catheter in about 1 min using a Harvard pump. Dynamic scanning started immediately after FBB injection, with 1.23-mm isotropic resolution using list-mode acquisition. During the PET imaging procedures, the participants rested quietly under dim illumination. To ensure that subjects did not fall asleep, they were monitored throughout the procedure and asked to keep their eyes open (no fixation cross). Information about head movement was collected using a cap with small light reflectors and a Polaris Vicra (Northern Digital) head-tracking system to minimize motion-related image blurring. Before FBB injection, a transmission scan was obtained using cesium-137 to correct for attenuation.

**FBB-PET Scans.** Each participant underwent two FBB-PET scans to measure ABB, one scan on a day following RW and another scan on a day following a night of SD. For this purpose, participants stayed overnight at the Clinical Center at the NIH before their scheduled SD or RW scans. For the SD condition, participants were instructed to wake up at 8:00 AM on the morning before the SD night. Upon arrival, participants were continuously accompanied by a nurse to ensure that they stayed awake for the SD condition during their stay. For the RW conditions, nurses observed whether patients were asleep every hour from 10:00 PM to 7:00 AM. On the day of RW, the participants were woken up at 7:00 AM. On both days, participants remained under supervision of the nurse and were brought to the PET imaging center before the scan, which started around 1:30 PM. Thus, participants remained awake for a total of about 31 h (including the scan length) during the SD condition. No food was given after midnight and caffeinated beverages were discontinued 24 h before the study. Patients had a light breakfast and lunch on the scan day. The order of RW and SD scans was counterbalanced across subjects and were, on average, 15 d apart (standard deviation = 19 d). In preparation for the scans, a catheter was placed for radiotracer injection. Dynamic scanning started right after intravenous injection of about 9.5 mCi (or less) of FBB, which lasted for 120 min. During FBB scans, participants were encouraged to listen to music to stay awake.

**PET Data Analysis.** FBB uptake in the brain was quantified using the SUVr using the whole cerebellum as reference region for images obtained from later time points (90–110 min) (57, 58). We chose the SUVr method given the recent evidence (for a radiotracer from the same family as FBB) that the SUVr approach of estimating tracer accumulation, relative to other non-invasive kinetic modeling approaches, had one of the highest correlations, with arterial input compartment modeling results ($R^2 = 0.95$), and had comparable accuracy, while maintaining much simpler modeling requirements and not suffering from voxel-level noise on model fitting (59). Despite the concern that SUVr estimates of radiotracer accumulation are affected by confounds such as blood-flow changes, recent work has suggested a limited influence of blood-flow changes on FBB SUVr measures obtained from the later time points, similar to the range used in our study (i.e., 90–110 min) (52). SUVr images for FBB were coregistered with individual subjects' T1-weighted images and resampled into 2-mm isotropic resolution before being transformed into the MNI space using the same normalization parameters that were generated for T1-weighted (and T2-weighted) images. For statistical parametric mapping, images were smoothed with a 4-mm kernel and masked at 0.5 SUVr to remove voxels outside the brain. We also performed a follow-up BPnd analysis to address the concern that FBB SUVr might have been affected by confounds, such as changes in blood flow and tracer clearance properties (52, 60). Specifically, regional BPnd was

calculated using a noninvasive simplified reference tissue model (SRTM) (61, 62) that was fitted to the time activity curve (0–120 min) derived from the subcortical cluster showing a significant effect of SD on FBB SUVr (Fig. 1*A*). We also corroborated SRTM-based BPnd findings using the reference Logan approach (Fig. S2) (59, 63). Kinetic modeling of dynamic PET data was performed in the PMOD Kinetic Modeling Tool v3.605 (PMOD Technologies). For consistency with the SUVr approach, the whole cerebellum was used as the reference region for the BPnd analyses.

**Pittsburgh Sleep Quality Index.** The PSQI questionnaire (64) was administered to all participants ($n = 20$). For the analyses, we used the number of SH and the TS quality score from the PSQI, because prior studies have linked these sleep measures to brain ABB (6, 65, 66). While SH is a self-report of participants' SH at night (excluding times spent awake in bed), TS is a composite of scores of sleep quality, latency, hours, efficiency, medication, disturbances because of one's health or sleep partners, and daytime life quality (lower TS indicates better sleep quality).

**APOE Genotyping.** SNPs of the APOE gene (rs7412 and rs429358) have been shown to influence brain glucose metabolism and ABB (67, 68). Accordingly, in all participants we genotyped rs7412 and rs429358 SNPs of APOE and computed a log of ORAD for each participant, following the methods of a previous study (69), normalized relative to the population risk for AD. We used ORAD to evaluate whether APOE influenced the association between sleep and ABB.

**Mood Questionnaires.** Mood questionnaires were collected periodically (five times) throughout the RW and SD scans in 18 of the 20 participants. Questionnaires were acquired at least an hour apart, prior (three measures), during (one measure), and after (one measure) the PET scans. On a scale of 0–10, subjects rated whether they felt alert, tired, hungry, friendly, happy, sad, anxious, irritable, social, confused, bored, comfortable, energetic, caffeine craving, and difficulty staying awake. The scores across the five time-points were averaged for each mood measure for each scan day. The first principle component of the standardized SD–RW difference scores for the 15 self-report mood measures was computed to summarize the change in mood from RW to SD into one component. This component accounted for 35.5% of the variance of the mood change measures (SD–RW) and was positively correlated with changes in ratings of feeling alert, friendly, happy, social, and energetic ($P < 0.01$, two-tailed), and negatively correlated with changes in ratings of feeling tired, anxious, and irritable from RW to SD ($P < 0.01$, two-tailed). Higher scores along this component reflected positive changes in mood from RW to SD.

**Statistical Parametric Mapping.** SPM8 (Wellcome Trust Centre for Neuroimaging) (70) was used for performing a voxelwise paired *t* test between SD and RW in FBB SUVr and voxelwise correlation between behavioral or genotyping measures and PET data. All effects were thresholded at $P \leq 0.015$ in SPM8 and a minimum cluster size of 300 voxels (2-mm isotropic). We chose this threshold based on the low sensitivity of FBB to soluble Aβ (28), but effects were corrected for multiple comparisons for cluster size using the random field theory (71) (family-wise error, FWE), or unless indicated, with false-discovery rate (FDR) in SPM8.

**Data Availability.** Subject-level measures used in the manuscript are available in Dataset S1. Please contact corresponding authors for additional information.

**ACKNOWLEDGMENTS.** We thank Joanna Fowler for the insightful comments; Christopher Wong, Lori Talagala, and Minoo McFarland for their assistance with behavioral and imaging data collection; David Goldman, Hui Sun, and Melanie Schwandt for assistance with APOE genotyping; Jeih-San Liow and Santi Bullich for insights about PET data analysis; and Kimberly Herman, Tom Lionetti, and Rosa Clark for assistance with monitoring participants. This work was supported by NIH/National Institute on Alcohol Abuse and Alcoholism intramural research program (Grant Y1AA3009).

1. Nedergaard M (2013) Neuroscience. Garbage truck of the brain. *Science* 340: 1529–1530.
2. Mestre H, Kostrikov S, Mehta RI, Nedergaard M (2017) Perivascular spaces, glymphatic dysfunction, and small vessel disease. *Clin Sci (Lond)* 131:2257–2274.
3. Xie L, et al. (2013) Sleep drives metabolite clearance from the adult brain. *Science* 342: 373–377.
4. Kang J-E, et al. (2009) Amyloid-β dynamics are regulated by orexin and the sleep-wake cycle. *Science* 326:1005–1007.
5. Tabuchi M, et al. (2015) Sleep interacts with aβ to modulate intrinsic neuronal excitability. *Curr Biol* 25:702–712.
6. Spira AP, et al. (2013) Self-reported sleep and β-amyloid deposition in community-dwelling older adults. *JAMA Neurol* 70:1537–1543.
7. Brown BM, et al.; AIBL Research Group (2016) The relationship between sleep quality and brain amyloid burden. *Sleep (Basel)* 39:1063–1068.
8. Sprecher KE, et al. (2015) Amyloid burden is associated with self-reported sleep in nondemented late middle-aged adults. *Neurobiol Aging* 36:2568–2576.
9. Ju Y-ES, et al. (2013) Sleep quality and preclinical Alzheimer disease. *JAMA Neurol* 70: 587–593.
10. Ju Y-ES, Lucey BP, Holtzman DM (2014) Sleep and Alzheimer disease pathology—A bidirectional relationship. *Nat Rev Neurol* 10:115–119.
11. Mucke L, Selkoe DJ (2012) Neurotoxicity of amyloid β-protein: Synaptic and network dysfunction. *Cold Spring Harb Perspect Med* 2:a006338.
12. Jagust W (2016) Is amyloid-β harmful to the brain? Insights from human imaging studies. *Brain* 139:23–30.
13. Spira AP, Gottesman RF (2017) Sleep disturbance: An emerging opportunity for Alzheimer's disease prevention? *Int Psychogeriatr* 29:529–531.
14. Lee H, et al. (2015) The effect of body posture on brain glymphatic transport. *J Neurosci* 35:11034–11044.
15. Iliff JJ, et al. (2012) A paravascular pathway facilitates CSF flow through the brain parenchyma and the clearance of interstitial solutes, including amyloid β. *Sci Transl Med* 4:147ra111.
16. Ooms S, et al. (2014) Effect of 1 night of total sleep deprivation on cerebrospinal fluid β-amyloid 42 in healthy middle-aged men: A randomized clinical trial. *JAMA Neurol* 71:971–977.
17. LaFerla FM, Green KN (2012) Animal models of Alzheimer disease. *Cold Spring Harb Perspect Med* 2:a006320.
18. Bu XL, et al. (October 31, 2017) Blood-derived amyloid-β protein induces Alzheimer's disease pathologies. *Mol Psychiatry*, 10.1038/mp.2017.204.
19. Deane R, Zlokovic BV (2007) Role of the blood-brain barrier in the pathogenesis of Alzheimer's disease. *Curr Alzheimer Res* 4:191–197.
20. Wei M, et al. (2017) Sleep deprivation induced plasma amyloid-β transport disturbance in healthy young adults. *J Alzheimers Dis* 57:899–906.
21. Lucey BP, et al. (December 8, 2017) Effect of sleep on overnight CSF amyloid-β kinetics. *Ann Neurol*, 10.1002/ana.25117.
22. Klunk WE, et al. (2004) Imaging brain amyloid in Alzheimer's disease with Pittsburgh compound-B. *Ann Neurol* 55:306–319.
23. Rowe CC, et al. (2008) Imaging of amyloid β in Alzheimer's disease with 18F-BAY94-9172, a novel PET tracer: Proof of mechanism. *Lancet Neurol* 7:129–135.
24. Fodero-Tavoletti MT, et al. (2012) In vitro characterization of [18F]-florbetaben, an Aβ imaging radiotracer. *Nucl Med Biol* 39:1042–1048.
25. Villemagne VL, et al. (2011) Amyloid imaging with (18)F-florbetaben in Alzheimer disease and other dementias. *J Nucl Med* 52:1210–1217.
26. Sabri O, Seibyl J, Rowe C, Barthel H (2015) Beta-amyloid imaging with florbetaben. *Clin Transl Imaging* 3:13–26.
27. Ni R, Gillberg P-G, Bergfors A, Marutle A, Nordberg A (2013) Amyloid tracers detect multiple binding sites in Alzheimer's disease brain tissue. *Brain* 136:2217–2227.
28. Yamin G, Teplow DB (2017) Pittsburgh compound-B (PiB) binds amyloid β-protein protofibrils. *J Neurochem* 140:210–215.
29. Schuff N, et al.; Alzheimer's Disease Neuroimaging Initiative (2009) MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. *Brain* 132:1067–1077.
30. De Santi S, et al. (2001) Hippocampal formation glucose metabolism and volume losses in MCI and AD. *Neurobiol Aging* 22:529–539.
31. Mander BA, Winer JR, Jagust WJ, Walker MP (2016) Sleep: A novel mechanistic pathway, biomarker, and treatment target in the pathology of Alzheimer's disease? *Trends Neurosci* 39:552–566.
32. Ju YS, et al. (2017) Slow wave sleep disruption increases cerebrospinal fluid amyloid-β levels. *Brain* 140:2104–2111.
33. Thal DR, Rüb U, Orantes M, Braak H (2002) Phases of A β-deposition in the human brain and its relevance for the development of AD. *Neurology* 58:1791–1800.
34. Aron L, Yankner BA (2016) Neurodegenerative disorders: Neural synchronization in Alzheimer's disease. *Nature* 540:207–208.
35. Bero AW, et al. (2011) Neuronal activity regulates the regional vulnerability to amyloid-β deposition. *Nat Neurosci* 14:750–756.
36. Cheng O, et al. (2015) Short-term sleep deprivation stimulates hippocampal neurogenesis in rats following global cerebral ischemia/reperfusion. *PLoS One* 10:e0125877.
37. Fernandes C, et al. (2015) Detrimental role of prolonged sleep deprivation on adult neurogenesis. *Front Cell Neurosci* 9:140.
38. Castellano JM, et al. (2011) Human apoE isoforms differentially regulate brain amyloid-β peptide clearance. *Sci Transl Med* 3:89ra57.
39. Jack CR, Jr, et al.; Alzheimer's Disease Neuroimaging Initiative (2010) Brain beta-amyloid measures and magnetic resonance imaging atrophy both predict time-to-progression from mild cognitive impairment to Alzheimer's disease. *Brain* 133: 3336–3348.
40. Vandenberghe R, et al. (2010) 18F-flutemetamol amyloid imaging in Alzheimer disease and mild cognitive impairment: A phase 2 trial. *Ann Neurol* 68:319–329.
41. Meyer-Luehmann M, et al. (2008) Rapid appearance and local toxicity of amyloid-β plaques in a mouse model of Alzheimer's disease. *Nature* 451:720–724.

42. Condello C, Schain A, Grutzendler J (2011) Multicolor time-stamp reveals the dynamics and toxicity of amyloid deposition. *Sci Rep* 1:19.
43. Rothman SM, Herdener N, Frankola KA, Mughal MR, Mattson MP (2013) Chronic mild sleep restriction accentuates contextual memory impairments, and accumulations of cortical Aβ and pTau in a mouse model of Alzheimer's disease. *Brain Res* 1529: 200–208.
44. Zhao HY, et al. (2017) Chronic sleep restriction induces cognitive deficits and cortical beta-amyloid deposition in mice via BACE1-antisense activation. *CNS Neurosci Ther* 23:233–240.
45. Camus V, et al. (2012) Using PET with 18F-AV-45 (florbetapir) to quantify brain amyloid load in a clinical environment. *Eur J Nucl Med Mol Imaging* 39:621–631.
46. Montagne A, et al. (2015) Blood-brain barrier breakdown in the aging human hippocampus. *Neuron* 85:296–302.
47. Cai Z, Hussain MD, Yan L-J (2014) Microglia, neuroinflammation, and beta-amyloid protein in Alzheimer's disease. *Int J Neurosci* 124:307–321.
48. Zhu B, et al. (2012) Sleep disturbance induces neuroinflammation and impairment of learning and memory. *Neurobiol Dis* 48:348–355.
49. Sehlin D, et al. (2016) Antibody-based PET imaging of amyloid beta in mouse models of Alzheimer's disease. *Nat Commun* 7:10759.
50. Wu JC, et al. (2006) Frontal lobe metabolic decreases with sleep deprivation not totally reversed by recovery sleep. *Neuropsychopharmacology* 31:2783–2792.
51. Ma N, Dinges DF, Basner M, Rao H (2015) How acute total sleep loss affects the attending brain: A meta-analysis of neuroimaging studies. *Sleep (Basel)* 38:233–240.
52. Bullich S, et al. (2017) Validation of non-invasive tracer kinetic analysis of 18F-florbetaben PET using a dual time-window acquisition protocol. *J Nucl Med* jnumed.117.200964.
53. Price JL, Drevets WC (2010) Neurocircuitry of mood disorders. *Neuropsychopharmacology* 35:192–216.
54. Mueller AD, Meerlo P, McGinty D, Mistlberger RE (2013) Sleep and adult neurogenesis: Implications for cognition and mood. *Sleep, Neuronal Plasticity and Brain Function*, eds Meerlo P, Benca RM, Abel T (Springer, Heidelberg, Germany), pp 151–181.
55. Dissel S, et al. (2017) Enhanced sleep reverses memory deficits and underlying pathology in *Drosophila* models of Alzheimer's disease. *Neurobiol Sleep Circadian Rhythms* 2:15–26.
56. Glasser MF, et al.; WU-Minn HCP Consortium (2013) The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* 80:105–124.
57. Catafau AM, et al. (2016) Cerebellar amyloid-β plaques: How frequent are they, and do they influence 18F-florbetaben SUV ratios? *J Nucl Med* 57:1740–1745.
58. Bullich S, et al. (2017) Optimal reference region to measure longitudinal amyloid-beta change with 18F-florbetaben PET. *J Nucl Med* jnumed.116.187351.
59. Heurling K, Buckley C, Van Laere K, Vandenberghe R, Lubberink M (2015) Parametric imaging and quantitative analysis of the PET amyloid ligand [(18)F]flutemetamol. *Neuroimage* 121:184–192.
60. Ossenkoppele R, Prins ND, van Berckel BN (2013) Amyloid imaging in clinical trials. *Alzheimers Res Ther* 5:36.
61. Gunn RN, Lammertsma AA, Hume SP, Cunningham VJ (1997) Parametric imaging of ligand-receptor binding in PET using a simplified reference region model. *Neuroimage* 6: 279–287.
62. Lammertsma AA, Hume SP (1996) Simplified reference tissue model for PET receptor studies. *Neuroimage* 4:153–158.
63. Logan J, et al. (1996) Distribution volume ratios without blood sampling from graphical analysis of PET data. *J Cereb Blood Flow Metab* 16:834–840.
64. Buysse DJ, Reynolds CF, 3rd, Monk TH, Berman SR, Kupfer DJ (1989) The Pittsburgh Sleep Quality Index: A new instrument for psychiatric practice and research. *Psychiatry Res* 28:193–213.
65. Choe YM, et al. (2016) Sleep quality in young and middle age-period is associated with cerebral amyloid burden in cognitively normal elderly people. *Alzheimers Dement* 12(Suppl):P171–P172.
66. Westwood AJ, et al. (2017) Prolonged sleep duration as a marker of early neurodegeneration predicting incident dementia. *Neurology* 88:1172–1179.
67. Liu C-C, Kanekiyo T, Xu H, Bu G (2013) Apolipoprotein E and Alzheimer disease: Risk, mechanisms and therapy. *Nat Rev Neurol* 9:106–118, and erratum (2013), 10.1038/nmeurol.2013.32.
68. Reiman EM, et al. (2005) Correlations between apolipoprotein E ε4 gene dose and brain-imaging measurements of regional hypometabolism. *Proc Natl Acad Sci USA* 102:8299–8302.
69. Corneveaux JJ, et al. (2010) Association of CR1, CLU and PICALM with Alzheimer's disease in a cohort of clinically characterized and neuropathologically verified individuals. *Hum Mol Genet* 19:3295–3301.
70. Friston KJ, et al. (1995) Analysis of fMRI time-series revisited. *Neuroimage* 2:45–53.
71. Poline J-B, Worsley KJ, Evans AC, Friston KJ (1997) Combining spatial extent and peak intensity to test for activations in functional imaging. *Neuroimage* 5:83–96.

pii: sp-00462-15 http://dx.doi.org/10.5665/sleep.5552

PEDIATRICS

# Cognitive Performance, Sleepiness, and Mood in Partially Sleep Deprived Adolescents: The Need for Sleep Study

June C. Lo, PhD; Ju Lynn Ong, PhD; Ruth L.F. Leong, BSSc; Joshua J. Gooley, PhD; Michael W.L. Chee, MBBS

*Centre for Cognitive Neuroscience, Neuroscience and Behavioral Disorders Program, Duke-NUS Medical School, Singapore*

**Study Objectives:** To investigate the effects of sleep restriction (7 nights of 5 h time in bed [TIB]) on cognitive performance, subjective sleepiness, and mood in adolescents.

**Methods:** A parallel-group design was adopted in the Need for Sleep Study. Fifty-six healthy adolescents (25 males, age = 15–19 y) who studied in top high schools and were not habitual short sleepers were randomly assigned to Sleep Restriction (SR) or Control groups. Participants underwent a 2-w protocol consisting of 3 baseline nights (TIB = 9 h), 7 nights of sleep opportunity manipulation (TIB = 5 h for the SR and 9 h for the control groups), and 3 nights of recovery sleep (TIB = 9 h) at a boarding school. A cognitive test battery was administered three times each day.

**Results:** During the manipulation period, the SR group demonstrated incremental deterioration in sustained attention, working memory and executive function, increase in subjective sleepiness, and decrease in positive mood. Subjective sleepiness and sustained attention did not return to baseline levels even after 2 recovery nights. In contrast, the control group maintained baseline levels of cognitive performance, subjective sleepiness, and mood throughout the study. Incremental improvement in speed of processing, as a result of repeated testing and learning, was observed in the control group but was attenuated in the sleep-restricted participants, who, despite two recovery sleep episodes, continued to perform worse than the control participants.

**Conclusions:** A week of partial sleep deprivation impairs a wide range of cognitive functions, subjective alertness, and mood even in high-performing high school adolescents. Some measures do not recover fully even after 2 nights of recovery sleep.

**Commentary:** A commentary on this article appears in this issue on page 497.

**Keywords:** adolescents, cognitive performance, mood, partial sleep deprivation, sleepiness

**Citation:** Lo JC, Ong JL, Leong RL, Gooley JJ, Chee MW. Cognitive performance, sleepiness, and mood in partially sleep deprived adolescents: the need for sleep study. *SLEEP* 2016;39(3):687–698.

**Significance**

Some of the world's most sleep deprived students live in East Asia where students excel in standardized academic tests. This might reinforce the notion that 'mind over matter' can overcome negative effects of chronic sleep restriction. We found that in adolescents, partial sleep deprivation of comparable duration and severity to that examined in studies on young healthy adults elicited equivalent or greater neurobehavioral deficits across several cognitive domains. Residual effects on sustained attention, speed of processing, and subjective alertness can still be observed even after 2 nights of recovery sleep. That even students from top high schools are susceptible to neurobehavioral deficits should cause policymakers and parents to reconsider if sleep should continue to be sacrificed for the sake of academic achievement.

## INTRODUCTION

Sleep curtailment in adolescents is a serious problem in many societies, but insufficient action is being taken to stem this tide. Approximately 75% of adolescents in the US[1] and more than 90% in Korea[2] and Japan[3] sleep less than the recommended 8–10 h a night.[4] Previously, the maturational delay in bedtime combined with early morning school were the principal reasons for shortened sleep in adolescence.[5] In recent years, increased electronic media use, higher homework load, and reduced parental control have contributed to further sleep curtailment in this age group.[6] In highly competitive societies in East Asia where voluntary sleep curtailment is most prevalent, there is widespread belief that greater effort and more time spent studying, perhaps at the expense of sleep, is mandatory for acceptable academic performance.[7] This viewpoint is sustained by the higher scores achieved on standardized tests by students from East Asian countries[8] who, on average, sleep 1 to 2 h less than their European[9,10] or Australian[10] counterparts. Although three decades of observational and experimental studies on sleep curtailment in adolescents have provided clear evidence for increased daytime sleepiness, the case for objective cognitive performance degradation following partial sleep deprivation has been less compelling,[6,11] prompting the current study.

### Effects of Partial Sleep Deprivation on Subjective Sleepiness and Mood

Sleep restricted adolescents have been consistently found to be more sleepy. Results of observational studies have revealed shorter sleep duration to be associated with higher levels of subjective sleepiness.[12,13] Moreover, experimental studies have shown that just 1 night of 4- to 5-h sleep opportunity reduces sleep latency in the Multiple Sleep Latency Test[14–16] and increases levels of subjective sleepiness.[15] After 5 nights of 6.5 h of time in bed (TIB), higher levels of subjective sleepiness have been corroborated by parental assessment.[17]

Short sleep duration has also been associated with greater emotional lability.[18] Compared to a well-rested condition, 2 nights of sleep restriction lowered self-reported positive affect.[19] Elevated negative affect ratings were observed after 5 nights of restriction to 6.5 h of TIB for sleep.[20]

### Cognitive Consequences of Partial Sleep Deprivation

In comparison to adults, the effects of shortened sleep on objectively measured cognitive performance in children and adolescents have been found to be relatively modest, leading some to suggest that adolescents may be more resistant to sleep loss.[21] Although several observational studies have found that speed of processing, sustained attention, working memory,

and executive function are poorer in children and adolescents who report shorter sleep,[22–24] other studies have failed to find a significant relationship between sleep duration and speed of processing,[23,25] working memory, or executive function.[12,23,25]

Experimental studies on the cognitive consequences of partial sleep deprivation in children and adolescents have yielded heterogeneous findings, possibly because of differences in the extent of partial sleep deprivation and the cognitive tasks used across studies. In relation to partial sleep deprivation, both the severity of sleep restriction each night and the number of nights sleep was restricted have generally been lesser than in adult studies. Most partial sleep deprivation studies in children and adolescents have either reduced TIB by only 1 h for a few nights[26] or have restricted sleep opportunity to 4 to 5 h for only 1 night.[14–16,27] Although partial sleep deprivation has been observed to impair attention,[26] working memory,[26] executive function,[16] and verbal creativity[16] in some studies, others have not found any significant decrement in attention,[14,15,27] executive function,[27] or speed of processing.[14,16,26]

Two studies investigated the cognitive effects of a longer period of sleep restriction. In one, 5 nights of sleep restricted to 6.5 h TIB resulted in increased student and parent reports of inattention, as well as problems with metacognition.[17] However, in a subset of these participants who underwent functional magnetic resonance imaging, the investigators found no objective deficit in working memory or executive function. These adolescent participants might have modulated task-related activation to mitigate any potentially deleterious effects of sleep restriction.[28]

In a second study,[21] participants restricted to 5, 6, 7, 8, or 9 h of TIB for 4 nights did not exhibit any deficit in attention, speed of processing, executive function, or working memory. Although total sleep time (TST) was reduced in each of the sleep-restricted groups, the duration of slow wave sleep was not affected, leading the investigators to propose that adolescents may be resilient to cognitive impairment following substantial sleep restriction because of the preservation of slow wave sleep.[21]

A recent meta-analysis[29] on both observational and experimental studies found that in school-age children, the correlation between short sleep duration and poor cognitive performance was very modest ($r = 0.08$). When various cognitive domains were analyzed separately, shorter sleep duration was only modestly associated with poorer executive function, and not at all with sustained attention – a cognitive domain highly sensitive to partial sleep deprivation in adults.[30,31]

In the current study, we evaluated the effect of 7 nights of partial sleep deprivation on adolescents, seeking to fill gaps left by previous studies. First, we recruited students from top high schools – the type of students many lay persons expect to transcend the need for sleep when motivated to attain desired goals. Second, the modest effects of partial sleep deprivation in prior experiments could have resulted from insufficiently severe sleep restriction compared to similar studies in adults. In addition, these milder degrees of sleep restrtiction are not representative of the sleep schedules encountered by students living in highly competitive societies. To examine this possibility, sleep was restricted to 5 h TIB for 7 consecutive nights. Third, to facilitate comparison with similar studies on adults, our test battery comprised tests commonly used in adults. An example is the Psychomotor Vigilance Task (PVT),[32] which is widely used in sleep deprivation studies on adults[33] but has not been used in studies on children and adolescents. Fourth, to enhance ecological validity of our findings, the current study was conducted in a dormitory instead of in a sleep laboratory. Although a natural setting was used, the instrumentation, tests, and test frequency were similar to those used in laboratory-based studies. In particular, sleep was evaluated using both actigraphy and polysomnography (PSG).

## METHODS

### Participants

Sixty participants were invited to participate in the Need for Sleep Study, a 2-w protocol aimed at characterizing changes in cognitive performance, subjective sleepiness, and mood associated with sleep curtailment in adolescents. Participants were between 15 and 19 y of age; had to have no history of any chronic medical condition, psychiatric illness, or sleep disorder; had a body mass index ≤ 30; were not habitual short sleepers (i.e. had an average actigraphically estimated TIB of < 6 h and no sign of sleep extension on weekends); had to consume fewer than five cups of caffeinated beverages a day; and must not have traveled across more than two time zones 1 mo prior to the experiment.

Participants were randomized into the sleep restriction (SR) and the control groups. They were not informed about their grouping until the first day of the 2-w protocol. Two participants withdrew several days prior to the study and one during the study for personal reasons. One participant did not comply with the experimental procedures and was excluded from all the analyses.

The resulting sample consisted of 56 participants (25 males, mean ± standard deviation of age = 16.6 ± 1.1 y). The SR (n = 30) and the control groups (n = 26) did not differ in age, sex distribution, body mass index, consumption of caffeinated beverages, nonverbal intelligence, levels of anxiety and depression, morningness-eveningness preference, levels of daytime sleepiness, symptoms of chronic sleep reduction, global score of the Pittsburgh Sleep Quality Index, and self-reported and actigraphically assessed sleep habits (Table 1; refer to the next section for details of screening instrumentation). Data from actigraphy during term time indicated that on weekdays, these participants slept less than the recommended 8–10 h,[4] and TIB and TST increased by more than 2 h from weekdays to weekends (Table 1).

### Recruitment and Screening

This study was approved by the Institutional Review Board of the National University of Singapore. Participants were recruited through sleep education talks in two high-ranking high schools (see endnote A), advertisements on the laboratory and social networking websites, as well as by word of mouth. All interested participants and their legal guardians were invited to attend a briefing session. Written informed consent was obtained from each participant and a legal guardian.

The Pittsburth Sleep Quality Index[34] was used to assess self-reported sleep timing, duration, and quality, whereas the

**Table 1**—Characteristics for the sleep restriction and the control groups.

| | Sleep Restriction Group | | Control Group | | | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | $t / \chi^2$ | P |
| n | 30 | – | 26 | – | – | – |
| Age (y) | 16.43 | 0.94 | 16.81 | 1.17 | 1.33 | 0.19 |
| Sex (% males) | 46.70 | – | 42.30 | – | 0.11 | 0.74 |
| Body mass index | 20.43 | 2.88 | 20.38 | 2.55 | 0.07 | 0.94 |
| Caffeinated drinks per day | 0.75 | 0.55 | 0.54 | 0.79 | 1.18 | 0.25 |
| Raven's Advanced Progressive Matrices score | 9.77 | 1.98 | 10.38 | 1.06 | 1.43 | 0.16 |
| Beck Anxiety Inventory score | 7.80 | 6.45 | 6.58 | 4.83 | 0.79 | 0.43 |
| Beck Depression Inventory score | 6.90 | 5.49 | 5.19 | 4.68 | 1.24 | 0.22 |
| Morningness-Eveningness Questionnaire score | 47.90 | 7.43 | 49.96 | 7.15 | 1.05 | 0.30 |
| Epworth Sleepiness Scale score | 7.77 | 3.59 | 6.19 | 3.57 | 1.64 | 0.11 |
| Chronic Sleep Reduction Questionnaire | | | | | | |
| Total score | 34.50 | 5.77 | 33.81 | 5.13 | 0.47 | 0.64 |
| Shortness of sleep | 12.37 | 2.39 | 12.50 | 2.30 | 0.21 | 0.83 |
| Irritation | 6.97 | 1.85 | 6.77 | 1.58 | 0.43 | 0.67 |
| Loss of energy | 7.43 | 1.94 | 7.00 | 1.65 | 0.89 | 0.38 |
| Sleepiness | 7.73 | 1.66 | 7.54 | 1.75 | 0.43 | 0.67 |
| Pittsburgh Sleep Quality Index | | | | | | |
| TIB on weekdays (h) | 6.12 | 1.03 | 5.94 | 1.14 | 0.63 | 0.54 |
| TIB on weekends (h) | 8.70 | 1.23 | 9.20 | 1.30 | 1.50 | 0.14 |
| TIB on average (h) | 6.86 | 0.87 | 6.87 | 0.87 | 0.07 | 0.95 |
| TST on weekdays (h) | 5.91 | 1.02 | 5.78 | 1.15 | 0.45 | 0.66 |
| TST on weekends (h) | 8.48 | 1.24 | 9.04 | 1.30 | 1.65 | 0.11 |
| TST on average (h) | 6.64 | 0.87 | 6.71 | 0.88 | 0.28 | 0.78 |
| Global score | 5.17 | 2.32 | 4.58 | 2.58 | 0.90 | 0.37 |
| Actigraphy | | | | | | |
| TIB on weekdays (h) | 6.40 | 0.94 | 6.09 | 0.85 | 1.24 | 0.22 |
| TIB on weekends (h) | 8.46 | 1.08 | 8.45 | 1.25 | 0.99 | 0.99 |
| TIB on average (h) | 6.98 | 0.72 | 6.76 | 0.77 | 1.08 | 0.29 |
| TST on weekdays (h) | 5.61 | 0.86 | 5.37 | 0.73 | 1.11 | 0.27 |
| TST on weekends (h) | 7.46 | 1.10 | 7.53 | 1.14 | 0.21 | 0.84 |
| TST on average (h) | 6.14 | 0.66 | 5.99 | 0.62 | 0.89 | 0.38 |
| Sleep efficiency (%) | 87.86 | 5.46 | 88.45 | 4.66 | 0.42 | 0.68 |

SD, standard deviation; TIB, time in bed; TST, total sleep time.

Morningness-Eveningness Questionnaire[35] evaluated morningness-eveningness preference. Participants completed the Chronic Sleep Reduction Questionnaire[36] to evaluate symptoms of chronic sleep restriction, the Epworth Sleepiness Scale[37] to examine levels of daytime sleepiness, and the Berlin Questionnaire[38] to screen for obstructive sleep apnea. The Beck Anxiety Inventory[39] and the Beck Depression Inventory[40] were used to probe for anxiety and depression respectively. Nonverbal intelligence was assessed using the Raven's Advanced Progressive Matrices.[41] Participants wore an actiwatch (Actiwatch 2, Philips Respironics, Inc., Pittsburg, PA) for 1 w during term time to evaluate sleep patterns. They also filled in a sleep diary during that week, which provided additional information for identifying bedtime and wake time on the actogram.

Each participant who met the inclusion criteria was interviewed by JCL or RLL to ensure they would be comfortable interacting with other participants and research staff, as well as living away from home during the 2-w study period.

### Two-Week Study Protocol

One week prior to the study, participants were required to adhere to a sleep-wake schedule that provided a 9-h nocturnal sleep opportunity (23:00–08:00). This was verified using wrist-worn actigraphy and was intended for circadian entrainment and for minimizing any effect of prior sleep restriction on sleep and cognitive performance.

The 2-w protocol (Figure 1A) was conducted in a boarding school after the school year had ended. In the first 3 nights (B1–B3), both SR and control participants had a 9 h nocturnal sleep opportunity (23:00–08:00) for adaptation and baseline characterization purposes. This was followed by a 7-night manipulation period (M1–M7) in which the SR group had 5 h (01:00–06:00) and the control group had 9 h (23:00–08:00) sleep opportunities. The protocol ended with 3 nights of 9-h recovery sleep (R1–R3: 23:00–08:00) for both groups.

All participants slept in twin-share, air-conditioned rooms, each with its own en-suite bathroom. Males and females were

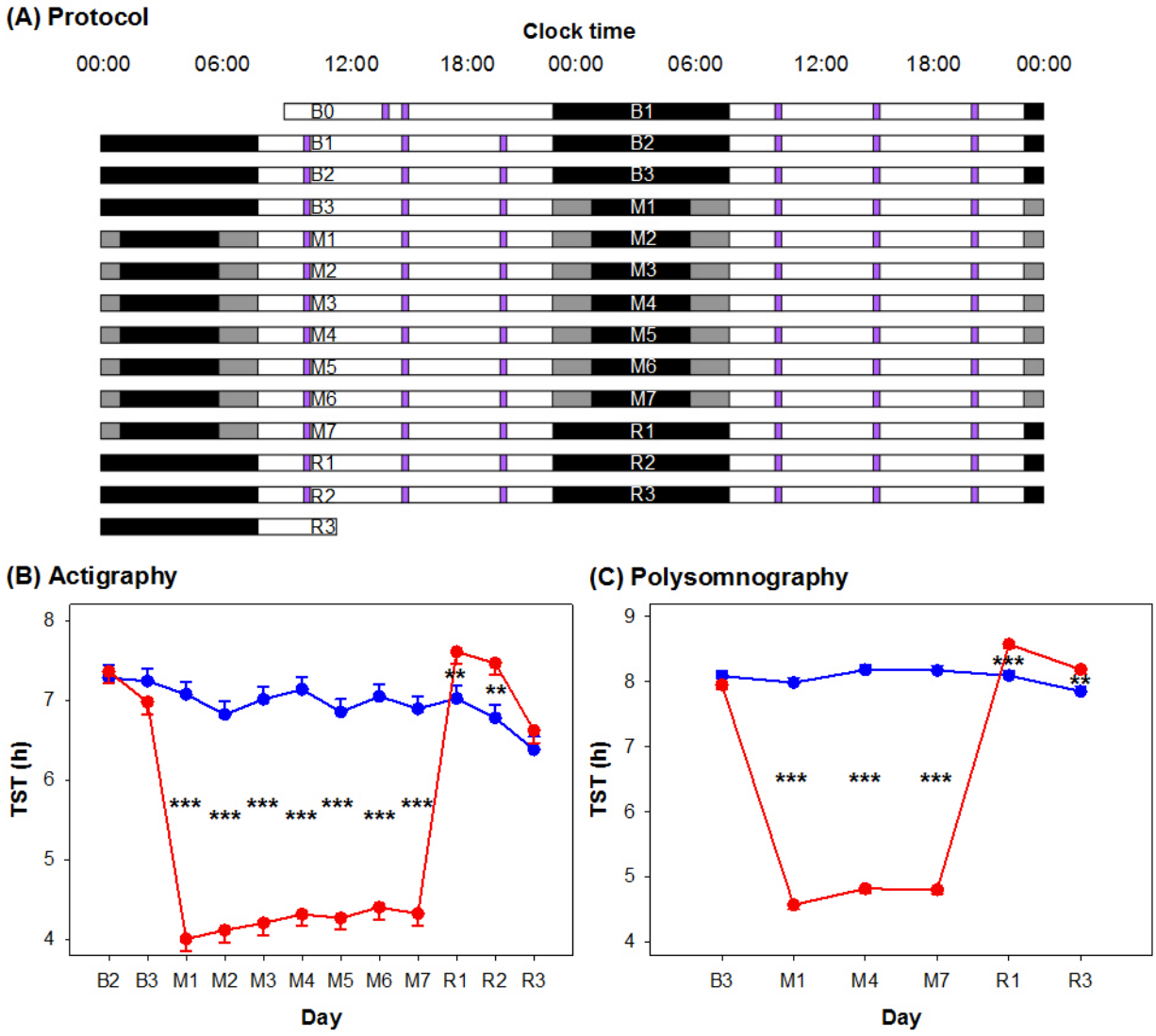


**Figure 1—(A)** Experimental protocol. The 2-w experimental protocol is illustrated in a double raster plot. Both the sleep restriction (SR) and the control groups had three adaptation and baseline nights (B1 to B3; time in bed [TIB] = 9 h), followed by 7 nights of sleep opportunity manipulation (M1 to M7; TIB = 5 h for SR [black bars] and 9 h for control [gray bars]), and 3 nights of recovery sleep (R1 to R3; TIB = 9 h). On most days, a cognitive performance test battery (purple bars) was administered at 10:00, 15:00, and 20:00. **(B)** Actigraphically and **(C)** polysomnographically assessed total sleep time (TST) of the SR (red lines) and the control (blue lines) groups from the adaptation and baseline period to the manipulation and recovery periods. Standard errors are illustrated. **P < 0.01; ***P < 0.001 for group contrasts.

housed in different buildings. The SR and the control groups were housed on different floors. Windows in each bedroom were fitted with blackout panels to prevent participants from being woken up by sunlight. Participants were provided with earplugs and allowed to adjust room temperature according to their own comfort. Apart from scheduled sleep periods, meal times, and cognitive testing periods, participants spent most of their time in a common room that received natural as well as artificial lighting. Participants were allowed to play board games, read, study, watch movies, and play games on their own electronic devices, in addition to interacting with research staff and other participants. Participants were under constant supervision of the research staff. Three main meals were served each day, and snacks were provided upon request. Caffeinated drinks, napping, and strenuous physical exercise were prohibited.

Sleep-wake patterns were continuously assessed with wrist-worn actigraphy, except for the first night, i.e. night B1, when all the actiwatches were charged. Each day, a computerized cognitive performance test battery was administered at 10:00, 15:00, and 20:00 (except for the first day [i.e. day B0]; Figure 1; see endnote B). Polysomnographic recordings were obtained on 7 nights: B1 and B3 for adaptation and baseline assessment, M1, M4, and M7 to monitor sleep changes from the beginning to the end of the manipulation period, and R1 and R3 for characterizing recovery sleep. Pulse oximetry was used on the first night to evaluate oxygen desaturations that might indicate undiagnosed obstructive sleep apnea. Here, we will report

findings regarding TST, whereas the sleep macrostructure and microstructure findings will be published separately.

### Cognitive Performance Test Battery

A computerized cognitive performance test battery was administered on 57 identical laptop computers (Acer Aspire E11, Acer Inc, Taipei, Taiwan). All tests were programmed in E-Prime 2.0 (Psychology Software Tools, Inc., Sharpsburg, PA). Each test battery lasted for approximately 25 min. Participants were required to wear earphones throughout the test battery to minimize distractions and for tone presentation during certain tasks. The test battery comprised 7 tasks presented in the following order: the Karolinska Sleepiness Scale (KSS)[42] the Sustained Attention to Response Task (SART),[43] the Symbol Digit Modalities Test (SDMT),[44] the verbal 1 and 3-back tasks,[31] the Mental Arithmetic Test (MAT),[45] the Positive and Negative Affect Scale (PANAS),[46] and the PVT.[32]

In the KSS,[42] participants rated their current level of subjective sleepiness using a nine-point Likert scale (1 very alert, 9 very sleepy, great effort to keep awake).

The SART[43] was used to assess sustained attention. Numbers ranging from 0 to 9 were presented sequentially on the screen for 250 msec, and participants were required to respond by pressing the spacebar on every trial, except when the target number '8' appeared. The target to non-target ratio was 15:85, and the inter-stimulus interval was fixed at 900 msec. Two nonparametric measures of sensitivity (*A'*) and response bias ($B''_D$) were used to quantify performance. *A'* is a measure of a participant's ability to discriminate between targets and non-targets, and is computed using the hit rate (number of non-target trials responded to × 100/85) and false alarm rate (number of target trials responded to × 100/15). *A'* ranges from 0 to 1, with 0.5 indicating performance at chance level. $B''_D$ is a measure of a participant's tendency toward liberal ($B''_D < 0$) or conservative ($B''_D > 0$) response behavior, where the former favors more responses and so is more likely to lead to responses when they are not required; the latter favors withholding responses, and as a result is less likely to result in false alarms when responses are not required. Neutrality is centered at 0 ($B''_D = 0$).

The two measures were derived with the following formula.[47,48]

$$\text{For } hit > fa,\ A' = \frac{1}{2} + \frac{(hit - fa) \times (1 + hit - fa)}{4 \times hit \times (1 - fa)}$$

$$\text{For } fa > hit,\ A' = \frac{1}{2} + \frac{(fa - hit) \times (1 + fa - hit)}{4 \times fa \times (1 - hit)}$$

$$B''_D = \frac{(1 - hit) \times (1 - fa) - (hit \times fa)}{(1 - hit) \times (1 - fa) + (hit \times fa)}$$

The SDMT[44] was used to measure speed of processing. In this task, participants were shown a key displaying nine pairs of digits and symbols. On every trial, a symbol appeared below the key, and participants were required to respond by inputting its corresponding digit (ranging from 1 to 9 on the keyboard) as quickly as possible. If participants did not respond in 15 sec, a beeping tone was presented until a response was recorded. This task lasted for 2 min. The total number of correct trials was used as the critical measure.

Verbal n-back tasks[31] were used to assess working memory and executive function. In this task, alphabets were presented sequentially for 1,000 msec with 3,000 msec inter-stimulus interval. Participants were required to decide whether the current stimulus matched with the one shown one (1-back) or three (3-back) items ago. The match to mismatch ratio was 8:24. We used the formulas stated above to derive measures of sensitivity (*A'*) and response bias ($B''_D$) to quantify performance.

The MAT[45] was used to measure speed of processing. This took the form of addition problems involving pairs of two-digit numbers that were shown on screen, and participants were required to solve them as quickly as possible. A beeping tone was presented if participants did not respond within 15 sec. The total number of correct trials in this 4-min task was used as the critical measure.

The PANAS[46] was used to assess positive and negative affect. Participants were shown 20 adjectives with 10 describing positive mood and 10 describing negative mood. Participants needed to respond using a five-point Likert scale (1 very slightly, 5 extremely).

A 10-min PVT[32] was used to measure levels of sustained attention. At random intervals varying from 2,000 msec to 10,000 msec, a counter on the computer screen started counting, and participants were required to respond as quickly as possible by pressing a key. A beeping tone was presented if no response was detected 10,000 msec after stimulus onset. The number of lapses (responses exceeding 500 msec) recorded during each PVT test was used as a measure of sustained attention.

### Actigraphy

Participants wore an actiwatch (Actiwatch 2, Philips Respironics, Inc., Pittsburgh, PA) around the wrist of their non-dominant hand during term time for screening purposes, during the 1-w pre-study period for verifying their compliance with the specified sleep schedule, as well as during the 2-w protocol. Data were collected at 2 min resolution and were scored with the Actiware software (version 6.0.2). TST was calculated using a medium sensitivity algorithm, with which an activity count greater than or equal to 40 was defined as waking. Participants also kept a sleep diary during the actigraphically monitored periods at home.

### Polysomnography

Electroencephalography (EEG) was performed using a SOMNOtouch recorder (SOMNOmedics GmbH, Randersacker, Germany) from two channels (C3 and C4 in the international 10–20 system) referenced to the contralateral mastoids. The common ground and reference electrodes were placed at Cz and FPz. Electrooculography (EOG) and submental electromyography (EMG) were also used. Impedance was kept below 5 kΩ for EEG electrodes and below 10 kΩ for EOG and EMG electrodes. Signals were sampled at 256 Hz and filtered between 0.2 and 35 Hz for EEG and between 0.2 and 10 Hz for EOG.

Sleep scoring analyses were performed using the FASST toolbox (http://www.montefiore.ulg.ac.be/~phillips/FASST.html).

EEG signals were band-pass filtered between 0.2 and 25 Hz. Scoring was performed visually by trained technicians following the criteria set by the American Academy of Sleep Medicine Manual for the Scoring of Sleep and Associated Events.[49]

### Statistical Analyses

Statistical analyses were performed with SAS 9.3 (SAS Institute, Cary, NC). We used a general linear mixed model with PROC MIXED to determine the effects of group, day (from day B3 to R2), and the group × day interaction on cognitive performance, sleepiness, and mood averaged across the three test batteries each day. We included performance on day B2 (see endnote C) as a covariate to control for group differences in baseline performance. To quantify the local effect size of partial sleep deprivation on each measure, we used a similar statistical model but excluded the recovery days to compute Cohen $f^2$ of the group × day interaction.[50] The cutoffs for small, medium, and large effect sizes were 0.02, 0.15, and 0.35, respectively.[51] We excluded data from the first five test batteries on days B0 and B1 in all the analyses to minimize influence of practice effects.

To assess the efficacy of our manipulation of sleep opportunities, we also used a general linear mixed model to determine the effects of group, day (from night B2 to R3 for actigraphic data, and from night B3 to R3 for PSG data), and group × day interaction on TST. PSG data from night B1, i.e., the adaptation night, was not included in the analysis. To ensure that the two groups followed the 9-h sleep schedule and did not differ in sleep duration the week prior to the 2-w protocol, we performed independent-samples *t* tests on actigraphically estimated TIB and TST.

## RESULTS

### Sleep Duration before and during the Protocol

One week before the 2-w protocol, both groups complied with the 9-h sleep schedule at home (mean ± standard error of the mean of TIB of the SR group: 8.78 ± 0.07 h versus control group: 8.84 ± 0.04 h, $t(53) = 0.68$, P = 0.50). There was no significant group difference in actigraphically estimated TST (SR: 6.89 ± 0.16 h versus control: 6.94 ± 0.11 h, $t(53) = 0.25$, P = 0.80), suggesting that sleep history did not differ between the two groups. The actigraphically estimated TST of 6.9 h appears short but is readily explained by actigraphy *underestimating* TST by approximately 1 h relative to PSG (see next section). As such, it is likely that our participants were well rested prior to the study.

We found that (1) the SR and the control groups had similar TST at baseline, (2) the partial sleep deprivation manipulation resulted in a large reduction in daily TST, and (3) the SR group had greater TST during the recovery nights. In the ensuing material, we provide a detailed breakdown of these points.

Actigraphy during the 2-w protocol revealed a significant group × day interaction on TST ($F(11,466) = 54.58$, P < 0.001). The two groups had similar actigraphically estimated TST on baseline nights (e.g., on B3, SR: 6.98 ± 0.15 h versus control: 7.24 ± 0.16 h, P = 0.23). During the manipulation period, TST was reduced to 4.01–4.41 h in the SR group and remained at approximately 7 h for the control group (Figure 1B). On the first 2 recovery nights, the SR group slept for 7.61 ± 0.15 h and 7.46 ± 0.15 h respectively, both longer than the last baseline night (P < 0.001 and P = 0.008) and significantly longer than the control group (R1: 7.02 ± 0.16 h, P = 0.01; R2: 6.78 ± 0.16 h, P = 0.002). On the third recovery night, TST of the SR group approached baseline level (P = 0.07), and the group difference disappeared (6.62 ± 0.16 h versus 6.38 ± 0.16 h, P = 0.23).

Actigraphy *underestimated* sleep duration by approximately 1 h relative to PSG. This systematic bias was less with higher sleep efficiency (i.e., as TST approached TIB) (Figure S1, supplemental material), and this was independent of the duration of sleep opportunity (TIB). Nevertheless, polysomnographic assessment of TST in response to sleep curtailment was, in general, congruent with the actigraphy findings (Figure 1C). The group × day interaction was statistically significant on TST ($F(5,179) = 572.14$, P < 0.001). TST in the last baseline night did not significantly differ between the two groups (SR: 7.95 ± 0.07 h versus control: 8.09 ± 0.07 h, P = 0.16). TST was maintained between 7.99 h and 8.18 h for the control group. The SR group showed a significant increase in TST from the beginning to the middle of the manipulation period (4.57 ± 0.07 h and 4.82 ± 0.07 h, P = 0.001). This was then maintained until the end of the sleep opportunity manipulation period (4.81 ± 0.07 h). In the first recovery night, not only was the SR group's TST significantly longer than the control group's (8.58 ± 0.07 h versus 8.09 ± 0.07 h, P < 0.001), it was significantly elevated from the baseline level (P < 0.001). On the third recovery night, TST of the SR group remained above baseline (P = 0.004) and significantly longer than the control (8.19 ± 0.07 versus 7.85 ± 0.08 h, P = 0.001).

### Effects of Partial Sleep Deprivation on Subjective Sleepiness, Cognitive Performance, and Mood

Cognitive performance, subjective sleepiness, and mood in the SR group were affected by partial sleep deprivation, as evidenced by a decrement in performance or reduced rate of improvement. Two nights of recovery sleep were insufficient to return performance to baseline levels on measures of sustained attention and subjective sleepiness (Figure 2). Although recovery sleep might have restored performance improvement in speed of processing tasks, performance of individuals with prior sleep restriction remained poorer than the well-rested control group. In general, the control group showed relatively stable cognitive performance, levels of subjective sleepiness, and mood throughout the protocol.

In evaluating sustained attention, we found a group × day interaction on the number of lapses in the PVT ($F(9,456) = 9.09$, P < 0.001). The SR group showed a monotonic increase in the number of lapses throughout the partial sleep deprivation period. The number of lapses was significantly reduced after the first recovery sleep episode (P < 0.001), but remained elevated relative to baseline after the first two nights of recovery sleep (P < 0.001; left panel in Figure 2A). Performance on the SART was less affected by partial sleep deprivation. Although the group × day interaction was also significant for *A'* in the SART ($F(9,457) = 2.02$, P = 0.04), a noticeable decrease in *A'* was found only after 4 nights of partial sleep deprivation.

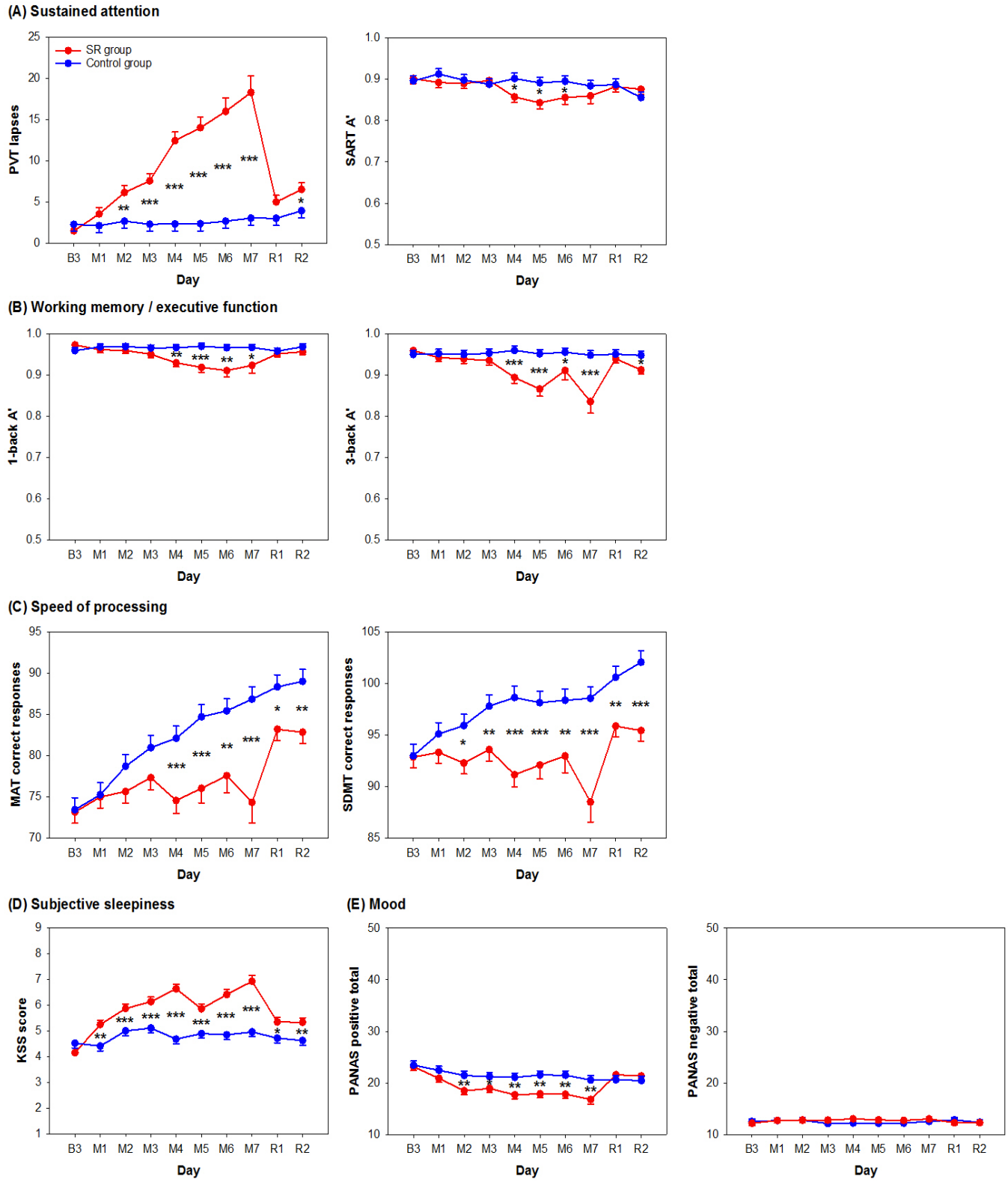


**Figure 2**—Effects of partial sleep deprivation on cognitive performance, subjective sleepiness, and mood. Daily average and standard errors of the sleep restriction (SR) (red lines) and the control (blue lines) groups from the days after the last baseline night (B3), after 1 to 7 nights of sleep manipulation (M1 to M7), and after 2 nights of recovery sleep (R1 and R2) were plotted for **(A)** sustained attention as indicated by the number of lapses in the Psychomotor Vigilance Task (PVT) and the sensitivity measure (*A'*) in the Sustained Attention to Response Task (SART), **(B)** working memory and executive functions as indicated by *A'* in the verbal 1- and 3-back tasks, **(C)** speed of processing as indicated by the number of correct responses in the Mental Arithmetic Test (MAT) and the Symbol Digit Modalities Test (SDMT), **(D)** subjective sleepiness level as indicated by score on the Karolinska Sleepiness Scale (KSS), and **(E)** positive and negative mood as indicated by the score on the Positive and Negative Affect Scale (PANAS). *P < 0.05; **P < 0.01; ***P < 0.001 for group contrasts.

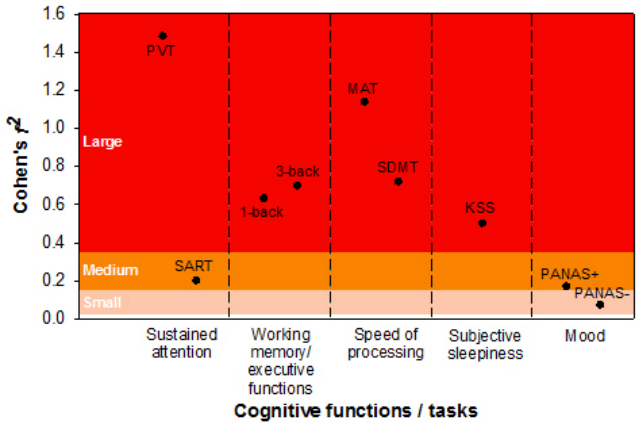


**Figure 3**—Effect size of partial sleep deprivation on cognitive performance, subjective sleepiness, and mood. Effect size is indicated by the local effect size (Cohen $f^2$) of group × day interaction on each cognitive measure (refer to the Methods section for further details). KSS, Karolinska Sleepiness Scale; MAT, Mental Arithmetic Test; PANAS, Positive and Negative Affect Scale (+, score on the positive affect subscale; −, score on the negative affect subscale); PVT, Psychomotor Vigilance Task; SART, Sustained Attention to Response Task; SDMT, Symbol Digit Modalities Test.

Performance returned to baseline levels after only 1 night of recovery sleep (P = 0.17; right panel in Figure 2A). This decrement in discriminability between targets and non-targets could not be explained by changes in response bias as $B''_D$ was not affected at all by partial sleep deprivation (group × day interaction: $F(9,457) = 1.30$, P = 0.23; Figure S2A, supplemental material). In terms of effect size, performance in the PVT was the most sensitive to partial sleep deprivation of all the tests in this study ($f^2 = 1.48$; Figure 3). In comparison, the *A'* in SART, was much less affected by sleep restriction ($f^2 = 0.20$; Figure 3).

In terms of working memory and executive function, we observed a significant group × day interaction on *A'* in the verbal 1-back task ($F(9,456) = 2.45$, P = 0.01). *A'* declined after 4 nights of sleep restriction and returned to baseline levels after one recovery sleep episode (P = 0.06; left panel of Figure 2B). The group × day interaction on $B''_D$ in the verbal 1-back task was not statistically significant ($F(9,457) = 1.70$, P = 0.09), and no significant group difference in the tendency toward conservative response behavior was observed throughout the protocol (left panel of Figure S2A). Hence, the decrement in *A'* in the SR group could not be accounted for by response bias. In the verbal 3-back task, there was also a significant group × day interaction on *A'* ($F(9,457) = 3.20$, P < 0.001). *A'* decreased after 4 nights of partial sleep deprivation and returned to baseline level after 1 night of recovery sleep (P = 0.15; right panel of Figure 2B) as observed in the verbal 1-back task. $B''_D$ in the verbal 3-back task did not reveal a statistically significant group × day interaction ($F(9,457) = 1.51$, P = 0.14; right panel of Figure S2B) and hence, could not explain the decrease in *A'* induced by sleep loss. The size of the partial sleep deprivation effect on 1- and 3-back was similar in magnitude ($f^2 = 0.63$ and 0.70; Figure 3), suggesting that cognitive decrement induced by partial sleep deprivation did not change with executive load.

In both the MAT and the SDMT, which are tests of speed of processing, performance improved with repeated testing, but this was attenuated in the SR group relative to the control group (group × day interaction for the MAT: $F(9,456) = 4.33$, P < 0.001; for the SDMT: $F(9,456) = 4.02$, P < 0.001). Interestingly, the largest improvement in both tasks was demonstrated by the SR group across the first recovery night (P < 0.001 for both tasks; Figure 2D). Nevertheless, after 2 nights of recovery sleep, performance of the SR group remained significantly poorer than the control group (P < 0.003 for both tasks). Although both speed of processing tasks revealed a similar pattern, performance in the MAT was a more sensitive measure of sleep loss than the SDMT ($f^2 = 1.14$ and 0.72; Figure 3).

Subjective sleepiness evaluated using the KSS showed a significant group × day interaction ($F(9,457) = 9.32$, P < 0.001). KSS score was elevated after only 1 night of partial sleep deprivation (P < 0.001), progressively increased thereafter, and plateaued toward the end of the manipulation period. Although KSS score in the SR group decreased after 1 night of recovery sleep (P < 0.001), it was still higher than the baseline value (P < 0.001) and at the level observed after 1 night of partial sleep deprivation (P = 0.61). This remained so even after the second recovery night (versus baseline: P < 0.001; versus M1: P = 0.65; Figure 2D). The effect of partial sleep deprivation on subjective sleepiness was in the medium range ($f^2 = 0.50$; Figure 3).

We found a significant group × day interaction on positive mood ($F(9,456) = 4.71$, P < 0.001). Positive mood decreased progressively during partial sleep deprivation, leveled off toward the end of the manipulation period, and returned to the baseline level after 1 night of recovery sleep (P = 0.36; left panel of Figure 2E). In contrast, the group × day interaction on negative mood was not statistically significant ($F(9,457) = 1.15$, P = 0.33). Negative mood appeared to stay at a low level throughout the protocol for both the SR and the control groups (main effect of day: $F(9,457) = 0.61$, P = 0.79; right panel of Figure 2E). Effect size measures showed that partial sleep deprivation had a medium effect on positive affect ($f^2 = 0.17$), but only a small effect on negative affect ($f^2 = 0.07$; Figure 3).

## DISCUSSION

Restricting adolescents' sleep to 5 h TIB for 7 nights led to cumulative degradation of sustained attention, working memory, executive function, and speed of processing. In contrast, the increase in subjective sleepiness and the reduction in positive mood leveled off in the course of the experiment. Residual effects on sustained attention and subjective sleepiness persisted after 2 nights of 9 h recovery sleep opportunity. Adolescents in the control condition consistently slept approximately 8 h each night, maintained their baseline levels of cognitive performance, subjective sleepiness, and mood, and even demonstrated improvement in speed of processing.

### Partial Sleep Deprivation Affects Even Academically Strong Students

Perhaps the most important finding of the current work is that even students from top schools who regularly sleep 2 to 3 h less than recommended for their age on weekday nights

experience significant neurobehavioral deficits when exposed to partial sleep deprivation.

Prior studies on adolescents with perhaps one exception[21] have used less harsh sleep restriction than that used here. Experimental studies of partial sleep deprivation in adults have used 3 to 6 h TIB for a minimum of 7 nights.[31,52,53] This constitutes about 2 to 4 h less actual sleep a night, assuming a norm of 7 to 8 h.[54] For adolescents, studies using 5 h TIB are in theory comparable to adult studies. A meta-analysis summarizing sleep duration data in the past century has shown that sleep in children and adolescents has decreased by about 75 min from the 20th century, with Asia showing one of the fastest rates of reduction.[55] This perhaps can be due to a disproportionately larger amount of time spent on school work in East Asia than in Western developed countries.[56] A survey in Korea involving a nationally representative sample of over 130,000 adolescents found that 43% reported sleeping less than 6 h each night.[2] Preliminary data from students of one of the feeder schools to the current work showed that on average, the actigraphically estimated TST was below 5.5 h during weekdays (unpublished data). As such, the severity and duration of sleep restriction used here has real-world relevance.

The severity of neurobehavioral deficits we observed in adolescents is comparable to if not greater than that observed with adults exposed to a similar degree of partial sleep deprivation. For example, young adults showed about 10 lapses in the PVT after 7 nights of 4 h TIB,[53] whereas an average of 18 lapses were found after 7 nights of 5 h TIB in the current study. Although many students seek to emulate elite performers who sleep little, the current data show that even students from the top performing country in a global test on reading, mathematics and science[8] are not spared and experience significant neurobehavioral deficits when undergoing partial sleep deprivation.

### Sustained Attention is the Most Affected Cognitive Domain, as in Adults

Previous studies have suggested that executive function and not attention is the cognitive domain most affected in partially sleep deprived children and adolescents.[11,16,57] These prior studies may not have found strong effects on attention because of differences in tests used to measure attention. The PVT is the most widely used test of sustained attention in adults. We detected monotonic deterioration in vigilance over the 7 nights of sleep restriction. The SART was less sensitive ($f^2 = 0.20$) than the PVT ($f^2 = 1.48$; Figure 3) highlighting the differential sensitivity to multi-night sleep restriction across tasks evaluating the same cognitive domain.

In terms of effect size, decline in speed of processing was the next most affected cognitive domain. This was slightly surprising given the absence of significant effects on this domain in prior studies on partial sleep deprivation in children and adolescents.[14,16,21,26] As in the case of sustained attention, we speculate that difference in task selection and severity of sleep restriction could explain these discrepancies. Although 1 night of recovery sleep might have restored the learning ability of the SR group, their performance failed to catch up with that of the control group. Whether additional nights of sufficient sleep can eliminate this group difference in performance remains to be investigated.

Working memory and executive function evaluated with 1- and 3-back tests were significantly affected by sleep restriction with an effect size of about half of that observed with the PVT. Interestingly, there was no additional decline in performance with increasing executive load (3 back versus 1 back). The absence of an incremental effect of load is similar to that observed with adults undergoing partial sleep deprivation[31] as well as visual short term memory and total sleep deprivation[58–60] and suggests that perceptual and attentional degradation[61] independently or together with maintenance failure[62] could underlie the performance decline attributed to executive function in the sleep deprived state.

### Two Nights of Recovery Sleep may not be Enough but Cognitive Domain Matters

Two nights of recovery sleep may not be sufficient to achieve a complete recovery in sustained attention, speed of processing, and alertness after 1 w of relatively severe sleep restriction in adolescents. These findings are reminiscent of a study where healthy young adults were restricted to 4 h TIB for 5 nights and the duration of recovery sleep was varied. Even 10 h of recovery sleep for a single night was insufficient to completely restore sustained attention to baseline levels, although speed of processing was restored.[63]

Particularly relevant to hard driving students, the residual effects of sleep deprivation may cumulate and subsequent exposure to sleep restriction following incomplete recovery may result in disproportional decline in performance (Dinges, unpublished data). Relevant to this point, the relative plateauing of subjective sleepiness compared to monotonically declining sustained attention and reduced improvement in speed of processing[53,64] could cause adolescents to underestimate the extent of their objective neurobehavioral deficit. Of particular concern in societies where sacrificing sleep for academic success is prevalent is that chronic fatigue becomes a new societal norm.[65]

### Poorer Positive Mood with Sleep Restriction

A decline in positive mood was observed after 2 nights of sleep restriction, similar to one previous study.[19] However, unlike another study,[20] we did not find an effect on negative mood. Although negative mood appeared to remain unaffected, many students remarked that the test items, e.g., guilty, scared, and afraid, were irrelevant to them. Sleep has been shown to modulate the processing of emotional memory.[66–68] Although this may have survival value, it could have negative effects on mental health. Indeed, a large behavioral risk factor survey found that shorter self-reported sleep duration in adolescents was associated with higher likelihood of reporting depressive symptoms and suicidal ideation.[2]

### Differences in Adolescent Sleep Assessed by Wrist Actigraphy and PSG

The Bland-Altman plot (Figure S1) indicates that when sleep efficiency was high (i.e., when TST approached TIB), there was overall good concordance between sleep duration measured by both actigraphy and PSG, but when sleep efficiency was low, there was a systematic *underestimation* of TST for actigraphy. The underlying reasons for the underestimation, rather

than overestimation,[69] of TST by actigraphy in our sample of adolescents and for the increased discrepancy between sleep duration assessed by PSG and wrist actigraphy as a function of sleep efficiency are unknown and remain to be investigated.

### Limitations and Future Studies

Sleep restriction was achieved by delaying bedtime and advancing wake time by 2 h so that the midpoints were aligned for both the sleep periods and the wake periods throughout the protocol to minimize circadian phase shifting. However, because test batteries were run at the same clock times, the duration of preceding wakefulness was always 2 h longer for the SR group during the manipulation period. The cognitive decrement associated with partial sleep deprivation might thus be accentuated by a longer duration of prior wakefulness before testing.

Our 7-night sleep restriction period was longer than the typical 5 study nights of 1 w when students curtailed their sleep. Although this might potentially limit the generalizability of our findings, it is not uncommon for highly competitive students to continue sleeping less than recommended on weekends in order to study. Furthermore, our finding that some cognitive functions failed to return to baseline levels after 2 nights of recovery sleep strongly signal the need to systematically evaluate the long-term effects of repeated cycles of sleep restriction and recovery on neurobehavioral deficits. Although we have unequivocally demonstrated neurobehavioral deficits using a standardized cognitive battery, the effect on ability to learn, to retain information, and to creatively reorganize learned material was not assessed. These higher-order cognitive functions are of critical interest and remain to be evaluated in future studies.

## CONCLUSION

Partial sleep deprivation in adolescents of comparable duration and severity to that examined in studies on young healthy adults elicited equivalent or greater neurobehavioral deficits across several cognitive domains. Residual effects on sustained attention, speed of processing, and subjective alertness can still be observed even after 2 nights of recovery sleep. That even students from top high schools are susceptible to neurobehavioral deficits following partial sleep deprivation should cause policymakers and parents to reconsider if sleep should continue to be sacrificed for the sake of academic achievement.

## ENDNOTES

A: Singapore was the top-ranked country out of 65 countries in the 2012 PISA examinations.[8] Most of our participants came from top ranked schools. All participants stayed in the boarding school during the 2-w protocol.

B: The label of day indicates the wake period after the corresponding sleep period. For example, day B2 refers to the day after the second baseline night, but before the third baseline night. This highlights the effect of sleep history on subsequent cognitive performance.

C: Performance on day B3 was not used as a covariate because these data were included in the effect of day in the statistical model. This model allows the evolution of cognitive performance, subjective sleepiness, and mood from the last baseline day (day B3) to be depicted.

## REFERENCES


1. National Sleep Foundation. Sleep in American poll: Teens and Sleep. Washington, DC: National Sleep Foundation, 2006.
2. Do YK, Shin E, Bautista MA, Foo K. The associations between self-reported sleep duration and adolescent health outcomes: what is the role of time spent on Internet use? Sleep Med 2013;14:195–200.
3. Ohida T, Osaki Y, Doi Y, et al. An epidemiologic study of self-reported sleep problems among Japanese adolescents. Sleep 2004;27:978–85.
4. Hirshkowitz M, Whiton K, Albert SM, et al. National Sleep Foundation's sleep time duration recommendations: methodology and results summary. Sleep Health 2015;1:40–3.
5. Carskadon MA. Sleep in adolescents: the perfect storm. Pediatr Clin North Am 2011;58:637–47.
6. Bryant NB, Gomez RL. The teen sleep loss epidemic: what can be done? Trans Issues Psychol Sci 2015;1:116–25.
7. Hsin A, Xie Y. Explaining Asian Americans' academic advantage over whites. Proc Natl Acad Sci U S A 2014;111:8416–21.
8. Organization for Economic Co-operation and Development. PISA 2012 results in focus: what 15-year-olds know and what they can do with what they know. 2014. Available from: http://www.oecd.org/pisa/keyfindings/pisa-2012-results-overview.pdf.
9. Gradisar M, Gardner G, Dohnt H. Recent worldwide sleep patterns and problems during adolescence: a review and meta-analysis of age, region, and sleep. Sleep Med 2011;12:110–8.
10. Olds T, Blunden S, Petkov J, Forchino F. The relationships between sex, age, geography and time in bed in adolescents a meta-analysis of data from 23 countries. Sleep Med Rev 2010;14:371–8.
11. Kopasz M, Loessl B, Hornyak M, et al. Sleep and memory in healthy children and adolescents - a critical review. Sleep Med Rev 2010;14:167–77.
12. Anderson B, Storfer-Isser A, Taylor HG, Rosen CL, Redline S. Associations of executive function with sleepiness and sleep duration in adolescents. Pediatrics 2009;123:e701–7.
13. Wolfson AR, Carskadon MA. Sleep schedules and daytime functioning in adolescents. Child Dev 1998;69:875–87.
14. Carskadon MA, Harvey K, Dement WC. Acute restriction of nocturnal sleep in children. Percept Mot Skills 1981;53:103–12.
15. Fallone G, Acebo C, Arnedt JT, Seifer R, Carskadon MA. Effects of acute sleep restriction on behavior, sustained attention, and response inhibition in children. Percept Mot Skills 2001;93:213–29.
16. Randazzo AC, Muehlbach MJ, Schweitzer PK, Walsh JK. Cognitive function following acute sleep restriction in children ages 10-14. Sleep 1998;21:861–8.
17. Beebe DW, Fallone G, Godiwala N, et al. Feasibility and behavioral effects of an at-home multi-night sleep restriction protocol for adolescents. J Child Psychol Psychiatry 2008;49:915–23.
18. Nixon GM, Thompson JM, Han DY, et al. Short sleep duration in middle childhood: risk factors and consequences. Sleep 2008;31:71–8.
19. Talbot LS, McGlinchey EL, Kaplan KA, Dahl RE, Harvey AG. Sleep deprivation in adolescents and adults: changes in affect. Emotion 2010;10:831–41.
20. Baum KT, Desai A, Field J, Miller LE, Rausch J, Beebe DW. Sleep restriction worsens mood and emotion regulation in adolescents. J Child Psychol Psychiatry 2014;55:180–90.
21. Voderholzer U, Piosczyk H, Holz J, et al. Sleep restriction over several days does not affect long-term recall of declarative and procedural memories in adolescents. Sleep Med 2011;12:170–8.
22. Gradisar M, Terrill G, Johnston A, Douglas P. Adolescent sleep and working memory performance. Sleep Biol Rhythms 2008;6:146–54.
23. Sadeh A, Gruber R, Raviv A. Sleep, neurobehavioral functioning, and behavior problems in school-age children. Child Dev 2002;73:405–17.

24. Steenari MR, Vuontela V, Paavonen EJ, Carlson S, Fjallberg M, Aronen E. Working memory and sleep in 6- to 13-year-old schoolchildren. J Am Acad Child Adolesc Psychiatry 2003;42:85–92.
25. Buckhalt JA, El-Sheikh M, Keller P. Children's sleep and cognitive functioning: race and socioeconomic status as moderators of effects. Child Dev 2007;78:213–31.
26. Sadeh A, Gruber R, Raviv A. The effects of sleep restriction and extension on school-age children: what a difference an hour makes. Child Dev 2003;74:444–55.
27. Kopasz M, Loessl B, Valerius G, et al. No persisting effect of partial sleep curtailment on cognitive performance and declarative memory recall in adolescents. J Sleep Res 2010;19:71–9.
28. Beebe DW, Difrancesco MW, Tlustos SJ, McNally KA, Holland SK. Preliminary fMRI findings in experimentally sleep-restricted adolescents engaged in a working memory task. Behav Brain Funct 2009;5:9.
29. Astill RG, Van der Heijden KB, Van Ijzendoorn MH, Van Someren EJ. Sleep, cognition, and behavioral problems in school-age children: a century of research meta-analyzed. Psychol Bull 2012;138:1109–38.
30. Basner M, Dinges DF. Maximizing sensitivity of the Psychomotor Vigilance Test (PVT) to sleep loss. Sleep 2011;34:581–91.
31. Lo JC, Groeger JA, Santhi N, et al. Effects of partial and acute total sleep deprivation on performance across cognitive domains, individuals and circadian phase. PLoS One 2012;7:e45987.
32. Dinges DF, Powell JW. Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. Beh Res Meth Instr Comp 1985;17:652–5.
33. Lim J, Dinges DF. Sleep deprivation and vigilant attention. Ann N Y Acad Sci 2008;1129:305–22.
34. Buysse DJ, Reynolds CF III, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res 1989;28:193–213.
35. Horne JA, Ostberg O. A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. Int J Chronobiol 1976;4:97–110.
36. Meijer AM. Chronic sleep reduction, functioning at school and school achievement in preadolescents. J Sleep Res 2008;17:395–405.
37. Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. Sleep 1991;14:540–5.
38. Netzer NC, Stoohs RA, Netzer CM, Clark K, Strohl KP. Using the Berlin Questionnaire to identify patients at risk for the sleep apnea syndrome. Ann Intern Med 1999;131:485–91.
39. Beck AT, Steer RA. Beck Anxiety Inventory Manual. San Antonio, TX: Harcourt Brace and Company, 1993.
40. Beck AT, Steer RA, Brown GK. Manual for the Beck Depression Inventory-II. San Antonio. TX: Psychological Corporation, 1996.
41. Raven J. Advanced progressive matrices: set II (1962 revision). London: H. K. Lewis, 1978.
42. Akerstedt T, Gillberg M. Subjective and objective sleepiness in the active individual. Int J Neurosci 1990;52:29–37.
43. Robertson IH, Manly T, Andrade J, Baddeley BT, Yiend J. 'Oops!': performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. Neuropsychologia 1997;35:747–58.
44. Smith A. Symbol Digit Modalities Test. Los Angeles, CA: Western Psychological Services, 1991.
45. Klein KE, Wegmann HM, Athanassenas G, Hohlweck H, Kuklinski P. Air operations and circadian performance rhythms. Aviat Space Env Med 1976;47:221–30.
46. Watson D, Clark LA, Tellegen A. Development and validation of brief measures of positive and negative affect: the PANAS scales. J Pers Soc Psychol 1988;54:1063–70.
47. Macmillan NA, Creelman CD. Detection theory: a user's guide. New York, NY: Cambridge University Press, 1991.
48. Pollack L, Norman DA. A non-parametric analysis of recognition experiments. Psychon Sci 1964;1:125–6.
49. Iber C, Ancoli-Israel S, Chesson A, Quan SF. The AASM manual for the scoring of sleep and associated events: rules, terminology, and technical specification, 1st ed. Westchester, IL: American Academy of Sleep Medicine, 2007.
50. Selya AS, Rose JS, Dierker LC, Hedeker D, Mermelstein RJ. A practical guide to calculating Cohen's f(2), a measure of local effect size, from PROC MIXED. Front Psychol 2012;3:111.
51. Cohen J. Statistical power analysis for the behavioral sciences, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
52. Belenky G, Wesensten NJ, Thorne DR, et al. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. J Sleep Res 2003;12:1–12.
53. Van Dongen HP, Maislin G, Mullington JM, Dinges DF. The cumulative cost of additional wakefulness: dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. Sleep 2003;26:117–26.
54. Steptoe A, Peacey V, Wardle J. Sleep duration and health in young adults. Arch Intern Med 2006;166:1689–92.
55. Matricciani L, Olds T, Petkov J. In search of lost sleep: secular trends in the sleep time of school-aged children and adolescents. Sleep Med Rev 2012;16:203–11.
56. Larson RW, Verma S. How children and adolescents spend time across the world: work, play, and developmental opportunities. Psychol Bull 1999;125:701–36.
57. Dahl RE. The impact of inadequate sleep on children's daytime cognitive function. Semin Pediatr Neurol 1996;3:44–50.
58. Chee MW, Chuah YM. Functional neuroimaging and behavioral correlates of capacity decline in visual short-term memory after sleep deprivation. Proc Natl Acad Sci U S A 2007;104:9487–92.
59. Chuah LY, Chee MW. Cholinergic augmentation modulates visual task performance in sleep-deprived young adults. J Neurosci 2008;28:11369–77.
60. Wee N, Asplund CL, Chee MW. Sleep deprivation accelerates delay-related loss of visual short-term memories without affecting precision. Sleep 2013;36:849–56.
61. Chee MW. Limitations on visual information processing in the sleep-deprived brain and their underlying mechanisms. Curr Opin Behav Sci 2015;1:56–63.
62. Tucker AM. Two independent sources of short term memory problems during sleep deprivation. Sleep 2013;36:815–7.
63. Banks S, Van Dongen HP, Maislin G, Dinges DF. Neurobehavioral dynamics following chronic sleep restriction: dose-response effects of one night for recovery. Sleep 2010;33:1013–26.
64. Rupp TL, Wesensten NJ, Balkin TJ. Trait-like vulnerability to total and partial sleep loss. Sleep 2012;35:1163–72.
65. Christakis NA, Fowler JH. Social contagion theory: examining dynamic social networks and human behavior. Stat Med 2013;32:556–77.
66. Stickgold R, Walker MP. Sleep-dependent memory triage: evolving generalization through selective processing. Nat Neurosci 2013;16:139–45.
67. Payne JD, Stickgold R, Swanberg K, Kensinger EA. Sleep preferentially enhances memory for emotional components of scenes. Psychol Sci 2008;19:781–8.
68. Sterpenich V, Albouy G, Boly M, et al. Sleep-related hippocampo-cortical interplay during emotional memory recollection. PLoS Biol 2007;5:e282.
69. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak CP. The role of actigraphy in the study of sleep and circadian rhythms. Sleep 2003;26:342–92.

## ACKNOWLEDGMENTS

The authors are grateful for Hans Van Dongen for his advice on statistical analyses, and Sher Yi Chiam and Benny Chin Seah Koh for their assistance in participant recruitment in their schools. We thank Amiya Patanaik and Jasmine Siudzinski for coding the cognitive tasks, and Jesisca Tandi, Wei Shan Cher, Pearlynne Chong, Bindiya Lakshmi Raghunath, V Vien Lee,

Shin Wee Chong, and Nicholas Ivan Chee for their effort in data collection and processing.

## SUBMISSION & CORRESPONDENCE INFORMATION

Submitted for publication August, 2015
Submitted in final revised form September, 2015
Accepted for publication October, 2015
Address correspondence to: Dr. Michael W.L. Chee, Centre for Cognitive Neuroscience, Duke-NUS Graduate Medical School, 8 College Road, Level 6, Singapore 169857; Tel: (+65) 6516 4916; Fax: (+65) 6221 8625; Email: michael.chee@duke-nus.edu.sg

## DISCLOSURE STATEMENT

This was not an industry supported study. Financial support was provided by the National Medical Research Council, Singapore (NMRC/STaR/0004/2008 and NMRC/STaR/015/2013) and The Far East Organization. The authors have indicated no financial conflicts of interest. This work was approved by the Institutional Review Board of the National University of Singapore (13-562). All participants provided written informed consent.

# Running enhances neurogenesis, learning, and long-term potentiation in mice

**Henriette van Praag*†, Brian R. Christie†‡, Terrence J. Sejnowski‡§, and Fred H. Gage*¶**

*Laboratory of Genetics and ‡Laboratory of Computational Neurobiology, Salk Institute for Biological Studies, La Jolla, CA 92037; and §Howard Hughes Medical Institute and Department of Biology, University of California at San Diego, La Jolla, CA 92093

Edited by Charles F. Stevens, Salk Institute for Biological Studies, La Jolla, CA, and approved September 2, 1999 (received for review July 20, 1999)

**Running increases neurogenesis in the dentate gyrus of the hippocampus, a brain structure that is important for memory function. Consequently, spatial learning and long-term potentiation (LTP) were tested in groups of mice housed either with a running wheel (runners) or under standard conditions (controls). Mice were injected with bromodeoxyuridine to label dividing cells and trained in the Morris water maze. LTP was studied in the dentate gyrus and area CA1 in hippocampal slices from these mice. Running improved water maze performance, increased bromodeoxyuridine-positive cell numbers, and selectively enhanced dentate gyrus LTP. Our results indicate that physical activity can regulate hippocampal neurogenesis, synaptic plasticity, and learning.**

New neurons are added continuously to certain areas of the adult brain, such as the hippocampus and olfactory bulb (1, 2). The functional significance of new hippocampal cells is not clear. In birds, food storage and retrieval experience correlate with changes in hippocampal size and neurogenesis (3). In mice, neurogenesis in the dentate gyrus increases with exposure to an enriched environment, and it is associated with improved learning (4). Similarly, voluntary physical activity in a running wheel enhances the number of new hippocampal cells (5). Although it is not known whether running also affects learning, it has been shown that physical activity facilitates recovery from injury (6) and improves cognitive function (7). Furthermore, trophic factors, associated with progenitor cell survival and differentiation (8), alterations in synaptic strength (9), long-term potentiation (LTP) (10), and memory function (11), are elevated after exercise (12, 13). At the cellular level, wheel running enhances the firing rate of hippocampal cells in a manner that correlates with running velocity (14). Thus, exercise may increase synaptic plasticity and learning, as well as neurogenesis. We designed the following experiments to test this hypothesis.

## Materials and Methods

**Subjects.** Thirty-four female C57BL/6 mice, 3 months old (The Jackson Laboratory) were divided into two groups of 17, the controls and the runners. The runners had free access to a running wheel equipped with an electronic counter. During the first 10 days animals received one 10-mg/ml intraperitoneal injection of 5-bromodeoxyuridine (BrdU; Sigma), dissolved in 0.9% NaCl, filtered sterile at 0.2 $\mu$m, at 50 $\mu$g/g of body weight per day to label dividing cells.

**Spatial Learning.** The mice were trained on a Morris water maze (15) with either two or four trials per day for 6 days. The platform was hidden 1 cm below the surface of water; it was made opaque with white nontoxic paint. The starting points were changed every day. Each trial lasted either until the mouse had found the platform or for a maximum of 40 s. At the end of each trial, the mice were allowed to rest on the platform for 10 s. The time to reach the platform (latency), the length of swim path, and the swim speed were recorded semi-automatically by a video tracking system (San Diego Instruments).

**Electrophysiology.** The animals were coded so that the experimenter was blind to the identity of individual mice. LTP experiments were done in the dentate gyrus and CA1 subfields of the hippocampus, both in the presence and in the absence of D-2-amino-5-phosphonovaleric acid (APV). The sequence of the experiments was randomized to prevent any possible order effects. Statistical analyses were carried out between pooled, rather than all, slices from individual mice. Data from all slices tested in the same condition from the same mouse were averaged to give a single value, whereas data from slices tested in different conditions from the same mouse were considered independent values.

Mice were anesthetized with fluorothane and decapitated, and the brains were quickly removed into chilled artificial cerebrospinal fluid (125.0 mM NaCl/2.5 mM KCl/1.25 mM $NaH_2PO_4$/25.0 mM $NaHCO_3$/2 mM $CaCl_2$/1.3 mM $MgCl_2$/10.0 mM dextrose/0.001 mM bicuculline methobromide, at pH 7.4), and continuously bubbled with 95% $O_2$/5% $CO_2$. One hemisphere was immediately placed in 4% paraformaldehyde in 0.1 M PBS and kept for histological analysis. The second hemisphere was affixed to a vibratome and cut into 400-$\mu$m slices. The slices were heated at 32°C for 30 min in a circulating perfusion chamber and then maintained at room temperature. Individual slices were transferred to the recording chamber as needed, and experiments were performed in artificial cerebrospinal fluid maintained at 32–34°C.

A sharpened tungsten, bipolar stimulating electrode and a 1-M$\Omega$ recording electrode filled with 3 mM KCl or 1 mM NaCl were used for testing LTP. For the dentate gyrus, electrodes were positioned in the middle third of the molecular layer, whereas for CA1 responses, a recording electrode and a stimulating electrode (toward CA3) were positioned in the stratum radiatum of field CA1, aided by a microscope (Olympus BX50wi) with a $\times$40 objective lens. Responses were evoked with single biphasic current pulses (10–400 $\mu$A), adjusted to yield a response $\approx$30% of maximum. All evoked responses were initially tested with paired-pulse stimuli (at 50, 100, 200, and 500 ms). For dentate recordings, only the responses that did not show paired-pulse facilitation were used to ensure that medial perforant path synapses were being examined (16, 17). Individual synaptic responses were elicited at 15-s intervals. After at least 10 min of stable baseline responses, LTP was induced by a burst of 50 pulses at 100 Hz; bursts were repeated four times at 30-s intervals. Recordings continued for 45 min after LTP induction.

**Immunohistochemistry.** Immunohistochemistry for BrdU and immunofluorescent triple labeling for BrdU, the neuronal marker

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: APV, D-2-amino-5-phosphonovaleric acid; BrdU, bromodeoxyuridine; LTP, long-term potentiation; NMDA, *N*-methyl-D-aspartate.

†H.v.P. and B.R.C. contributed equally to this work.

¶To whom reprint requests should be addressed. E-mail: fgage@salk.edu

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

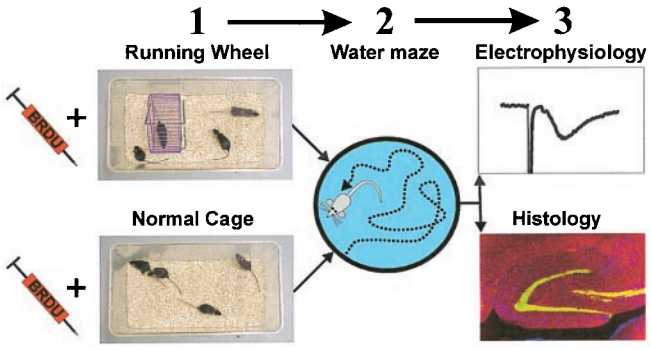


**Fig. 1.** Flowchart of the experiment. Controls were housed in standard 30- by 18-cm cages, whereas runners had 48- by 26-cm housing with free access to a running wheel (1). Mice in both conditions received BrdU (50 $\mu$g/g per day) injections for the first 10 days after their housing assignment. After 1 month in their respective environments, mice were tested in the water maze (2) between days 30 and 36 or between days 43 and 49. Mice were anesthetized and decapitated between days 54 and 118; one hemisphere was used for electrophysiology; the other was used for immunocytochemistry (3).

NeuN, and the glial marker calcium-binding protein S100$\beta$ were performed on free-floating 40-$\mu$m coronal sections that were pretreated by denaturing DNA, as described previously (5). The antibodies were mouse anti-BrdU (Boehringer Mannheim), 1:400; rat anti-BrdU ascites fluid (Accurate, Harlan Sera-Lab, Loughborough, England) for triple labeling, 1:100; rabbit anti-S100$\beta$ (Swant, Bellinoza, Switzerland) 1:2500; and mouse anti-NeuN (kindly provided by R. J. Mullen, University of Utah), 1:20. To determine the number of BrdU-labeled cells, we stained for BrdU by using the peroxidase method (ABC system, with biotinylated donkey anti-mouse IgG antibodies and diaminobenzidine as chromogen; Vector Laboratories). The fluorescent secondary antibodies used were FITC-labeled anti-mouse IgG, Texas red-labeled anti-rat IgG, and Cy5-labeled anti-rabbit IgG (Jackson ImmunoResearch), 6 $\mu$l/ml.

**Stereology.** BrdU-positive cells were counted in a one-in-six series of sections (240 $\mu$m apart) through a $\times$40 objective (Leitz) throughout the rostro caudal extent of the granule cell layer. A one-in-six series of adjacent sections stained with 0.5 mg/ml Hoechst 33342 in Tris-buffered saline (Molecular Probes) for 15 min was used to measure granule cell layer volume. We used a semiautomatic stereology system (StereoInvestigator, MicroBrightfield) and a $\times$10 objective to trace the granule cell area. The granule cell reference volume was determined by summing the traced granule cell areas for each section multiplied by the distance between sections sampled. The number of BrdU-labeled cells was then related to granule cell layer sectional volume and multiplied by the reference volume to estimate total number of BrdU-positive cells.

## Results

Mice were assigned to either control ($n$ = 17) or runner ($n$ = 17) conditions. Mice in the runner group ran an average distance of 4.78 $\pm$ 0.41 kilometers per day. During the first 10 days, animals received one intraperitoneal BrdU injection per day to label dividing cells. Thereafter, animals continued in their respective experimental conditions for 2 to 4 months. Mice were tested on the water maze task between day 30 and day 49. Between day 54 and day 118, mice were anesthetized with fluorothane and decapitated. One half of the brain was used for electrophysiological experiments. The other half was kept for immunocytochemistry (Fig. 1).

To assess spatial learning, mice were tested in the Morris water maze over 6 days. Mice were trained with four trials per day

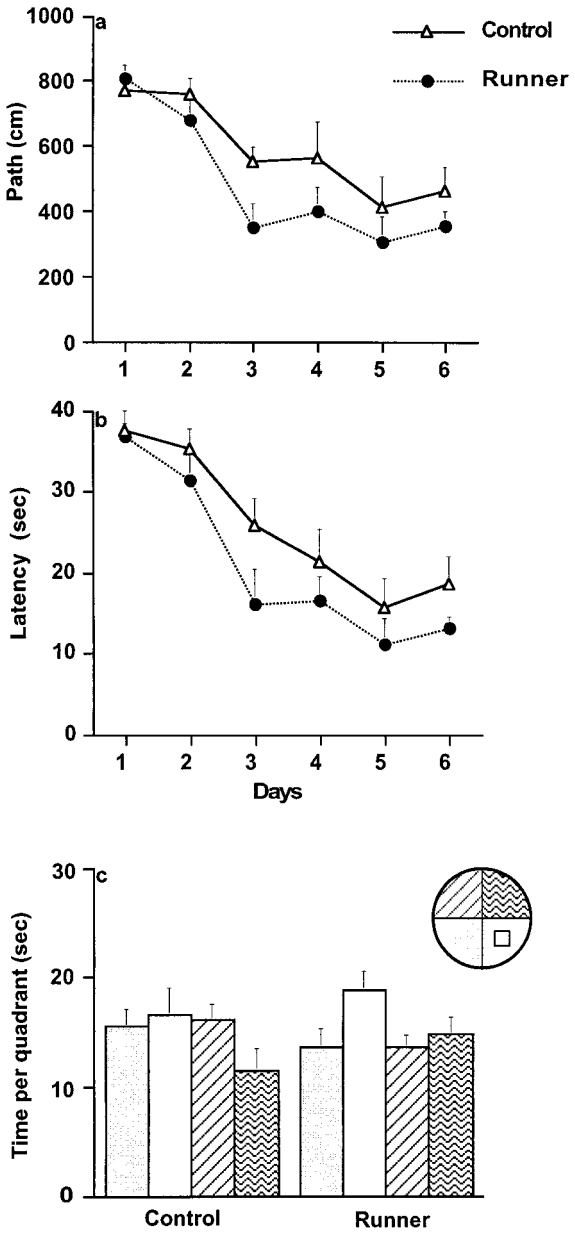


**Fig. 2.** Water maze learning in controls and runners trained with two trials per day (four-trial data are not shown here, but see description in *Results*). Mice were trained over 6 days to find the hidden platform in the Morris water maze. A significant difference developed between the groups ($P < 0.04$) in path length (*a*) and ($P < 0.047$) in latency (*b*). Results of probe test 4 hr after the last trial on day 6 (*c*).

(controls, $n$ = 8, and runners, $n$ = 8) between days 30 and 36, or two trials daily (controls, $n$ = 9 and runners, $n$ = 9) between days 43 and 49. When mice were trained with four trials per day, ANOVA with repeated measures (days) showed no difference between the groups in path length ($F_{(1,14)} = 0.56$, $P > 0.47$), latency ($F_{(1,14)} = 0.08$, $P > 0.79$), or swim speed ($F_{(1,14)} = 1.5$, $P > 0.24$). However, when mice were trained by using the more challenging two-trials-per-day paradigm, acquisition of the task was significantly better in the runners than in the controls, showing decreased path length ($F_{(1,16)} = 4.99$, $P < 0.04$; Fig. 2*a*) and latency ($F_{(1,16)} = 4.61$, $P < 0.047$; Fig. 2*b*) to the platform. These results were not confounded by swim speed, because there was no significant difference between the groups in this regard

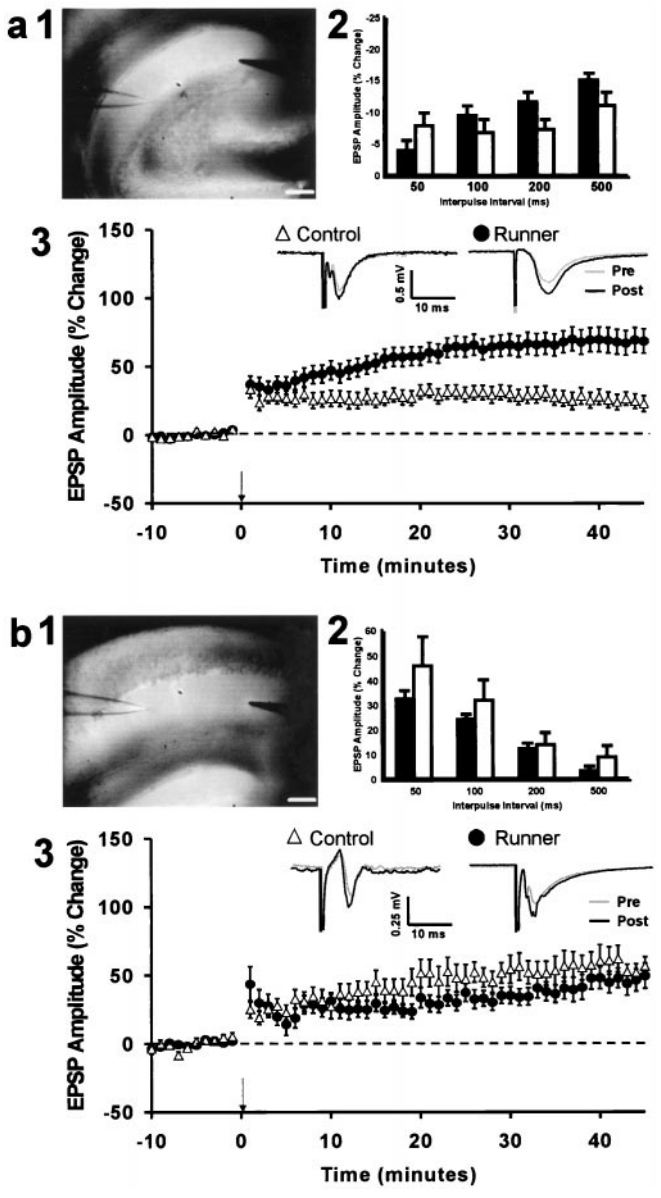


**Fig. 3.** LTP in dentate gyrus (*a*) and area CA1 (*b*). (*a1, b1*) Digital images of hippocampal slices showing the position of the stimulation and recording electrodes. (*a2, b2*) Paired-pulse facilitation at 50-, 100-, 200-, and 500-ms interpulse intervals. EPSP, excitatory postsynaptic potential. There was no difference between slices from controls and runners ($P > 0.91$). (*a3, b3*) Time course of LTP in slices from controls (△) and runners (●) . In addition, representative examples are shown of evoked responses immediately before (Pre) and 30 min after (Post) induction of LTP. Example waveforms are the average of 20 responses recorded over a 5-min period. Population spikes were apparent in some animals in each group after LTP induction. (Scale bars under *a1* and *b1* indicate 250 $\mu$m.)

($F_{(1,16)} = 0.59$, $P > 0.29$). To test retention of the task, the platform was removed for a 60-s probe test 4 hr after the last trial on day 6. Mice that had been trained with four trials per day spent more time in the platform quadrant than in all others (controls, $F_{(3,28)} = 57.17$, $P < 0.0001$; runners, $F_{(3,28)} = 16.70$, $P < 0.0001$). Mice trained with two trials per day did not show a significant preference for the platform quadrant (controls, $F_{(3,32)} = 1.53$, $P > 0.23$; runners, $F_{(3,32)} = 2.56$, $P < 0.07$), although for runners, the trend appeared stronger (Fig. 2*c*). Taken together, our results indicate that running enhances acquisition on the water maze task.

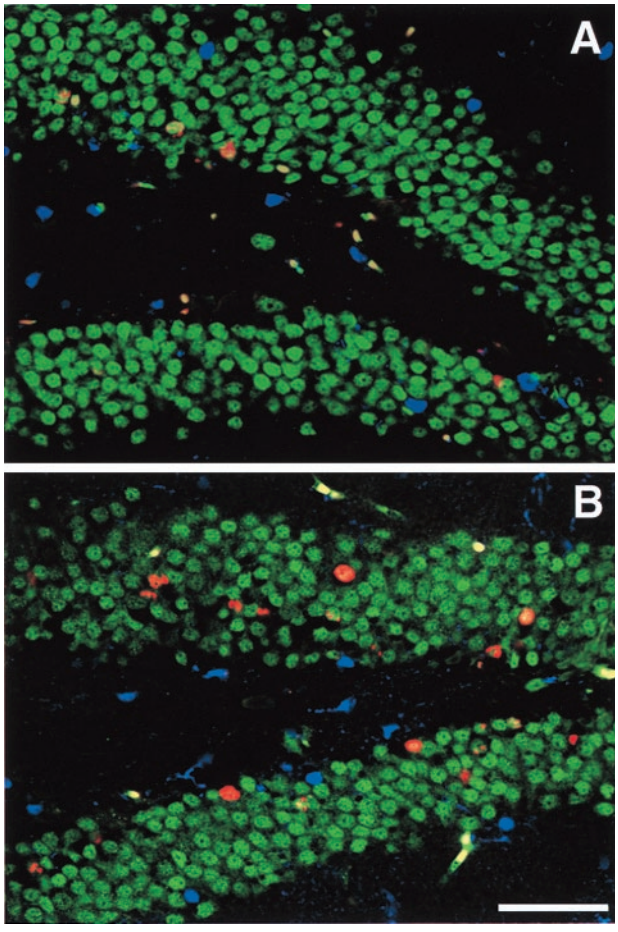


**Fig. 4.** Confocal images of BrdU-positive cells in control (*A*) and runner coronal sections (*B*). Sections were immunofluorescent triple-labeled for BrdU (red), NeuN, indicating neuronal phenotype (green), and S100$\beta$, selective for glial phenotype (blue). (Scale bar indicates 50 $\mu$m.)

Exercise can affect steroid hormone and stress levels, which in turn could influence learning, LTP, and neurogenesis (18–20). Therefore, after completion of behavioral testing, blood samples were collected retro-orbitally (controls, $n = 8$; runners, $n = 9$) on day 53 at 14:00 hr under isoflurane (4%) anesthesia. Radioimmunoassay for plasma corticosterone was performed by using the manufacturer's protocol for a commercial kit (ICN). No differences between the groups were observed (controls, 59 ± 13.1 ng/ml; runners, 49.4 ± 7.7 ng/ml; $t_{(15)} = 0.65$, $P > 0.52$). However, the possibility remains that running induces noncognitive, affective variables that could influence behavior.

To determine whether running affects synaptic plasticity, LTP was studied in hippocampal slices from the same 5- to 7-month-old controls and runners that had received BrdU and had been tested in the water maze. Recordings were made in the dentate gyrus, because this is where running-induced changes in cell proliferation and survival occur (5), and in area CA1. For dentate recordings only medial perforant path synapses were examined (refs. 16 and 17; Fig. 3*a2*). No differences were found between the groups in the initial excitatory postsynaptic potential (EPSP) amplitudes, suggesting no effect of running on basal synaptic efficacy (dentate gyrus: controls, 0.39 ± 0.04 mV, runners, 0.42 ± 0.05 mV ($t_{(22)} = 0.45$, $P > 0.66$); CA1: controls, 0.35 ± 0.07 mV, runners, 0.32 ± 0.08 mV ($t_{(12)} = 0.29$, $P > 0.78$)). The recordings in the dentate gyrus showed that EPSP amplitude was significantly increased 45 min after the administration of high-frequency stimuli (controls, 17 slices from 10 mice ($t_{(9)}$

NEUROBIOLOGY

**Table 1. BrdU-positive cell number and phenotype**

| Group | No. with BrdU label | % NeuN$^+$ | % S100$\beta^+$ | % Neither | Volume, mm$^3$ |
|---|---|---|---|---|---|
| Control | 1,409 ± 132 | 81.5 ± 2.4 | 5.5 ± 1.9 | 13.5 ± 3.2 | 0.438 |
| Runner | 3,746 ± 800* | 92.8 ± 1.7* | 1.8 ± 0.8 | 5.5 ± 1.6* | 0.481 |

Controls and runners received BrdU (50 $\mu$g/g per day) from day 1 to day 10. Survival of BrdU-labeled cells was determined 2 to 4 months after the last BrdU injection. Runners ($n = 8$), as compared to controls ($n = 8$), had significantly more BrdU-positive cells ($t_{(14)} = 2.88$, $P < 0.012$). The percentages of cells double-labeled for NeuN were greater in runners than in controls ($t_{(14)} = 3.82$, $P < 0.002$), and a lower percentage of BrdU-positive cells that were labeled for neither S100$\beta$ nor NeuN was observed in runners than in controls ($t_{(14)} = 2.26$, $P < 0.04$). Means are ± SEM. *, Significantly different from controls.

$= 2.34$, $P < 0.047$); runners, 20 slices from 14 mice ($t_{(13)} = 3.82$, $P < 0.002$); paired $t$ test)). The groups were significantly different from each other ($F_{(1,22)} = 4.66$, $P < 0.042$), demonstrating that running increased dentate gyrus LTP (Fig. 3*a3*). Robust LTP was also elicited in the CA1 Schaffer collateral pathway (controls, 12 slices from 7 mice ($t_{(6)} = 3.59$, $P < 0.011$); runners, 7 slices from 7 mice ($t_{(6)} = 2.55$, $P < 0.044$); paired $t$ test). However, there was no difference in the magnitude of CA1 LTP between controls and runners ($F_{(1,12)} = 0.22$, $P > 0.65$), (Fig. 3*b3*).

LTP in the medial perforant path-to-dentate granule cells and in Schaffer collateral CA1 synapses has been shown to depend on the activation of *N*-methyl-D-aspartate (NMDA) receptors (21). To determine whether LTP in runners was mediated by NMDA receptors, 50 $\mu$M NMDA receptor antagonist APV was added to the bath at the onset of recordings. APV completely blocked the induction of LTP, as demonstrated by the lack of an increase of percent baseline amplitude from values before induction [dentate gyrus: controls, four slices from four mice ($t_{(3)} = 0.51$, $P > 0.65$), runners, 11 slices from six mice ($t_{(5)} = 0.27$, $P > 0.79$); CA1: controls, five slices from five mice ($t_{(4)} = 0.44$, $P > 0.68$), runners, five slices from five mice ($t_{(4)} = 0.26$, $P > 0.81$); paired $t$ test].

BrdU-positive cells were analyzed in the remaining hemisphere of animals used for electrophysiology. The total number of BrdU-labeled cells was significantly greater in runners than in controls. In addition, 50 cells per animal were analyzed by confocal microscopy (Zeiss, Bio-Rad) for coexpression of BrdU and NeuN for neuronal phenotype and S100$\beta$ for glial phenotype. Runners had a significantly higher percentage of BrdU-positive cells that colabeled for NeuN. The groups did not differ with regard to astrocytic fate of the newborn cells (Fig. 4; Table 1).

## Discussion

Running enhances neurogenesis, water maze performance, and LTP in medial perforant path-to-dentate gyrus synapses. Spatial navigation in running mice was improved as compared with controls when mice were trained with two trials rather than four trials daily. Research by others on the effects of forced treadmill exercise on water maze learning in C57BL/6 mice also showed no difference with four trials daily, whereas tasks that are more complex demonstrated that exercising mice perform better (7). It is possible that increased neurogenesis in the runners contributes to learning. Indeed, several factors that elevate production of new neurons are also associated with enhanced learning. Both running and living in an enriched environment double the number of surviving newborn cells (5) and improve water maze performance (4). In addition, survival of cells born prior to spatial training may be enhanced by hippocampal-dependent learning (22). Moreover, treatment with hormones, such as estrogen, increases cell proliferation (23) and improves memory function (24). In contrast, factors that reduce neurogenesis, such as corticosterone treatment, stress, and aging, are associated with diminished performance on spatial learning tasks (20, 25). In our present study, corticosterone levels were similar in controls and runners; however, it is possible that other steroid hormone levels changed and influenced neurogenesis and learning.

The characteristics of LTP, including rapid formation, stability, synapse specificity, and reversibility, make it an attractive model for certain forms of learning and memory (26). Basic mechanisms underlying LTP in the dentate gyrus and CA1 subfields appear unchanged in runners and controls. Induction of LTP was completely blocked by the NMDA receptor antagonist APV (21). In addition, paired-pulse facilitation characteristics were similar to those in previous reports (16, 17), suggesting no running-induced alterations in presynaptic transmitter release. However, running did induce a specific increase in dentate gyrus LTP, concurrent with enhanced neurogenesis and improved water maze performance, raising the possibility that newborn granule cells play a role in increased dentate gyrus LTP. Although newborn cells are a small percentage of the total cells in the granule cell layer, it is possible that new neurons affect hippocampal physiology more than do mature cells (27). Indeed, in immature rats, dentate gyrus LTP lasts longer than in adults (28). Newborn cells in the adult brain may be similar to those generated during development.

The hypothesis that these new cells may mediate increased synaptic plasticity and improved learning is an attractive one; however, several other variables may be important as well. Increased exercise may result in elevated trophic factor production (12, 13), angiogenesis (29), and serotonin levels (30). Some of these variables may exert multiple effects, influencing learning, LTP, and neurogenesis. For example, increased serotonin levels enhance cell proliferation [B. L. Jacobs, P. Tanapat, A. J. Reeves, and E. Gould (1998) *Soc. Neurosci. Abstr.* **24**, 796.4]. Reduction of serotonin by 5,7-dihydroxytryptamine lesions diminishes dentate LTP (31). In addition, mutation of 5-HT$_{2c}$ receptors impairs medial perforant path-to-dentate gyrus, but not CA1, LTP and learning (32). It remains to be determined whether such factors induce running-enhanced synaptic plasticity and learning independently, or act indirectly, supporting survival and connectivity of newborn neurons.

Taken together, our findings demonstrate that voluntary running results in long-lasting survival of BrdU-positive cells, improved performance in a water maze, and a selective increase in dentate gyrus LTP. Further research will determine whether causal links can be found among these correlated measurements.

We thank Mary Lynn Gage and Barry Jacobs for comments on the manuscript, Linda Kitabayashi for assistance with confocal imaging, and Deborah Christie for graphic design. We also thank Alice Smith, Tony Slimp, Karen Suter, and coworkers in the Salk Institute Animal Research Facility for their support. This work was funded by the National Institute on Aging, the National Institute of Neurological Disorders and Stroke, the Lookout Fund, the Pasarow Foundation, the Holfelder Foundation, and the American Paralysis Association.

1. Gage, F. H., Kempermann, G., Palmer, T. D., Peterson, D. A. & Ray, J. (1998) *J. Neurobiol.* **36,** 249–266.
2. Goldman, S. A. & Luskin, M. B. (1998) *Trends Neurosci.* **21**, 107–114.
3. Clayton, N. S. & Krebs, J. R. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 7410–7414.
4. Kempermann, G., Kuhn, H. G. & Gage, F. H. (1997) *Nature (London)* **386,** 493–495.
5. van Praag, H., Kempermann, G. & Gage, F. H. (1999) *Nat. Neurosci.* **2,** 266–270.
6. Johansson, B. B. & Ohlsson, A. (1997) *Exp. Neurol.* **139**, 322–327.
7. Fordyce, D. E. & Wehner, J. M. (1993) *Brain Res.* **619**, 111–119.
8. Ray, J., Palmer, T. D., Suhonen, J., Takahashi, J. & Gage, F. H. (1997) in *Isolation, Characterization, and Utilization of CNS Stem Cells,* eds. Christen, Y. & Gage, F. H. (Springer, Berlin), pp. 129–149.
9. Schuman, E. M. (1999) *Curr. Opin. Neurobiol.* **9,** 105–109.
10. Patterson, S. L., Grover, L. M., Schwartzkroin, P. A. & Bothwell, M. (1992) *Neuron* **9,** 1081–1088.
11. Fischer, W., Wictorin, K., Bjorklund, A., Williams, L. R., Varon, S. & Gage, F. H. (1987) *Nature (London)* **329,** 65–68.
12. Neeper, S. A., Gomez-Pinilla, F., Choi, J. & Cotman, C. (1995) *Nature (London)* **373**, 109.
13. Gomez-Pinilla, F., So, V. & Kesslak, J. P. (1998) *Neuroscience* **85,** 53–61.
14. Czurko, A., Hirase, H., Csicsvari, J. & Buzsaki, G. (1999) *Eur. J. Neurosci.* **11**, 344–352.
15. Morris, R. G. M. (1984) *J. Neurosci. Methods* **11,** 47–60.
16. McNaughton, B. L. (1980) *Brain Res.* **199,** 1–19.
17. Colino, A. & Malenka, R. C. (1993) *J. Neurophysiol.* **69,** 1150–1159.
18. Tharp, G. D. (1975) *Med. Sci. Sports Exercise* **7,** 6–11.
19. Shors, T. J., Seib, T. B., Levine, S. & Thompson, R. F. (1989) *Science* **244,** 224–226.
20. McEwen, B. S. (1999) *Annu. Rev. Neurosci.* **22,** 105–122.
21. Wigstrom, H. & Gustafsson, B. (1986) *J. Physiol.* **81,** 228–236.
22. Gould, E., Beylin, A., Tanapat, P., Reeves, A. & Shors, T. J. (1999) *Nat. Neurosci.* **2,** 260–265.
23. Tanapat, P., Hastings, N. B., Reeves, A. J. & Gould, E. (1999) *J. Neurosci.* **19,** 5792–5801.
24. Luine, V. N., Richards, S. T., Wu, V. Y. & Beck, K. D. (1998) *Horm. Behav.* **34,** 149–162.
25. Krugers, H. J., Douma, B. R., Andringa, G., Bohus, B., Korf, J. & Luiten, P. G. (1997) *Hippocampus* **7,** 427–436.
26. Bliss, T. V. P & Collingridge, G. L. (1993) *Nature (London)* **361,** 31–39.
27. Gould, E., Tanapat, P., Hastings, N. B. & Shors, T. J. (1999) *Trends Cogn. Sci.* **3,** 186–192.
28. Bronzino, J. D., Abu-Hasballah, K., Austin La France, R. J. & Morgane, P. J. (1994) *Hippocampus* **4,** 439–446.
29. Isaacs, K. R., Anderson, B. J., Alcantara, A. A., Black, J. E. & Greenough, W. T. (1992) *J. Cereb. Blood Flow Metab.* **12,** 110–119.
30. Chaouloff, F. (1997) *Med. Sci. Sports Exercise* **29,** 58–62.
31. Bliss, T. V., Goddard, G. V. & Riives, M. (1983) *J. Physiol.* **334,** 475–491.
32. Tecott, L. H., Logue, S. F., Wehner, J. M. & Kauer, J. A. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 15026–15031.

# High impact running improves learning

**ARTICLE** *in* NEUROBIOLOGY OF LEARNING AND MEMORY · MAY 2007

Impact Factor: 3.65 · DOI: 10.1016/j.nlm.2006.11.003 · Source: PubMed



**11 AUTHORS**, INCLUDING:

Manfred Fobker
Universitätsklinikum Münster
**94** PUBLICATIONS **3,755** CITATIONS
SEE PROFILE

Karsten Krüger
Justus-Liebig-Universität Gießen
**60** PUBLICATIONS **634** CITATIONS
SEE PROFILE

Catharina Korsukewitz
University of Münster
**21** PUBLICATIONS **289** CITATIONS
SEE PROFILE

Stefan Knecht
Universitätsklinikum Düsseldorf
**252** PUBLICATIONS **8,512** CITATIONS
SEE PROFILE

Available online at www.sciencedirect.com
ScienceDirect
Neurobiology of Learning and Memory 87 (2007) 597–609
Neurobiology of Learning and Memory
www.elsevier.com/locate/ynlme

# High impact running improves learning

Bernward Winter [a,*,1], Caterina Breitenstein [a,b,1], Frank C. Mooren [c], Klaus Voelker [d], Manfred Fobker [e], Anja Lechtermann [d], Karsten Krueger [c], Albert Fromme [d], Catharina Korsukewitz [a], Agnes Floel [a], Stefan Knecht [a,b]

[a] *Department of Neurology, University of Muenster, Muenster, Germany*
[b] *IZKF Muenster, University of Muenster, Muenster, Germany*
[c] *Institute of Sports Medicine, Justus-Liebig University of Giessen, Giessen, Germany*
[d] *Institute of Sports Medicine, University Hospital of Muenster, Muenster, Germany*
[e] *Institute of Clinical Chemistry and Laboratory Medicine, University Hospital of Muenster, Muenster, Germany*

Received 12 September 2006; revised 30 October 2006; accepted 6 November 2006
Available online 20 December 2006

**Abstract**

Regular physical exercise improves cognitive functions and lowers the risk for age-related cognitive decline. Since little is known about the nature and the timing of the underlying mechanisms, we probed whether exercise also has *immediate* beneficial effects on cognition. Learning performance was assessed directly after high impact anaerobic sprints, low impact aerobic running, or a period of rest in 27 healthy subjects in a randomized cross-over design. Dependent variables comprised learning speed as well as immediate (1 week) and long-term (>8 months) overall success in acquiring a novel vocabulary. Peripheral levels of brain-derived neurotrophic factor (BDNF) and catecholamines (dopamine, epinephrine, norepinephrine) were assessed prior to and after the interventions as well as after learning. We found that vocabulary learning was 20 percent faster after intense physical exercise as compared to the other two conditions. This condition also elicited the strongest increases in BDNF and catecholamine levels. More sustained BDNF levels during learning after intense exercise were related to better short-term learning success, whereas absolute dopamine and epinephrine levels were related to better intermediate (dopamine) and long-term (epinephrine) retentions of the novel vocabulary. Thus, BDNF and two of the catecholamines seem to be mediators by which physical exercise improves learning.


*Keywords:* Learning; Memory consolidation; Catecholamines; Dopamine; Epinephrine; Brain-derived neurotrophic factor; Physical exercise; Language; Arousal

## 1. Introduction

Physical exercise seems to be beneficial to cognition. Epidemiological studies show that more frequent (self-reported) regular physical activity is associated with a reduced risk for age-related neurodegenerative diseases, like dementia or Parkinson's disease (Abbott et al., 2004; Colcombe et al., 2004; Larson et al., 2006; Laurin, Verreault, Lindsay, MacPherson, & Rockwood, 2001; van Gelder et al., 2004; Weuve et al., 2004). Beneficial effects of exercise on cognition may, however, be due to an overall healthier life style (non-smoking, better nutrition) in already cognitively high functioning subjects (Abbott et al., 2004; Kalmijn et al., 2000). Another confounder is that a preexisting, yet undiagnosed cognitive disorder may have led to a concomitant reduction in physical activity (Weuve et al., 2004). Thus, longitudinal intervention studies are better suited to determine the link between physical exercise and cognition. These studies show that several months of regular physical exercise led to improved mental functions or a slower cognitive decline in elderly subjects (Colcombe & Kramer, 2003 for a meta-analysis), with varying effect sizes for different cognitive functions (Fig. 1).

* Corresponding author. Fax: +49 251 83 48181.
*E-mail address:* bwinter@uni-muenster.de (B. Winter).
[1] These authors contributed equally.

1074-7427/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.nlm.2006.11.003

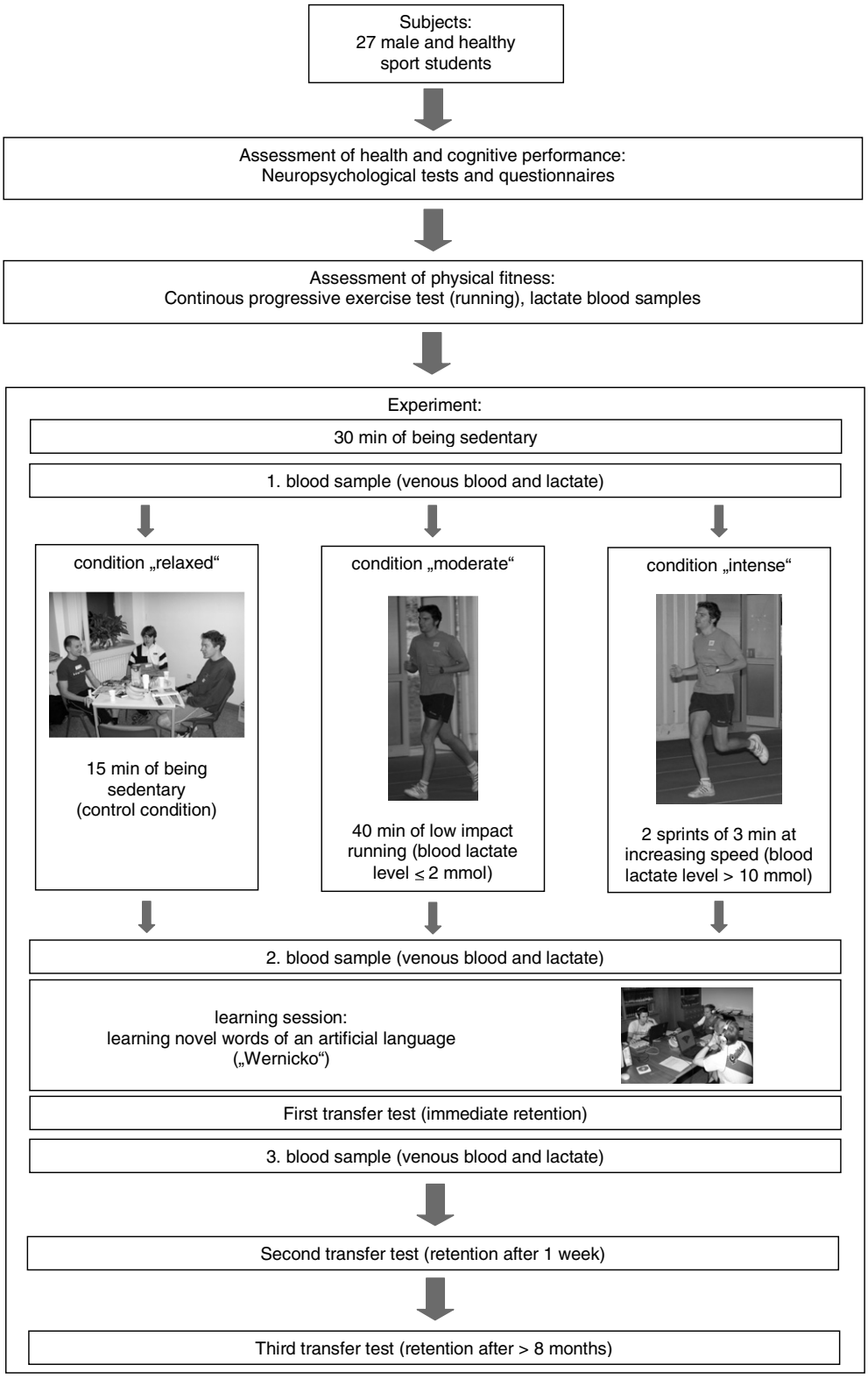


Fig. 1. Design of the present cross-over study showing the different interventions, points of measurement and retention.

Several studies probed the effect of acute bouts of exercise on cognitive functions (Etnier et al., 1997; Tomporowski, 2003; Tomporowski & Ellis, 1986 for a review). Most of these studies, however, did not directly assess effects on learning or memory, but rather investigated the effect of exercise on various neuropsychological measures, like simple reaction time tasks (e.g., Hogervorst, Riedel, Jeukendrup, & Jolles, 1996), or on exercise-related tasks like decision making in soccer (e.g., McMorris & Graydon, 1997). Studies probing the effect of exercise on memory led

to divergent results, depending on the length and intensity of the exercise intervention. After short-duration anaerobic exercise (up to 2 min), short-term memory was facilitated (Davey, 1973). During or immediately after long anaerobic exercise (5–40 min), no effects on memory were found (Sjoberg, 1980; Tomporowski, Ellis, & Stephens, 1987). When the exercise condition led to dehydration, either no effects or even negative effects on memory were noted (Cian, Barraud, Melin, & Raphel, 2001; Cian et al., 2000).

Similar results were obtained for healthy children: Sibley and Etnier reported positive effects of regular physical exercise for various cognitive tasks, but not for "pure" memory performance in their recent meta-analysis (Sibley & Etnier, 2003). Findings in children with attention deficit disorder are less conclusive: positive effects of exercise on disturbing behaviors have been found (Allison, Faith, & Franklin, 1995; Tantillo, Kesick, Hynd, & Dishman, 2002), but effects on cognition have not been reported so far (Craft, 1983).

In contrast to the preliminary evidence regarding the relation of physical exercise and cognition in humans, animals consistently showed improved learning after daily physical exercise for up to 7 months (Anderson et al., 2000; Baruch, Swain, & Helmstetter, 2004; Fordyce & Farrar, 1991; van Praag, Christie, Sejnowski, & Gage, 1999). Negative reports can be explained by an interaction of task complexity and heterogeneous task performance of the animals (Braszko, Kaminski, Hryszko, Jedynak, & Brzosko, 2001) or may be due to the use of a non-voluntary exercise condition (forced treadmill running; Burghardt, Fulk, Hand, & Wilson, 2004).

In animal studies, upregulations of various neurotransmitters in the brain, especially dopamine and norepinephrine, were found (Hattori, Naoi, & Nishino, 1994; Meeusen & De Meirleir, 1995; Sutoo & Akiyama, 2003). In addition to catecholamines, the release of neurotrophic factors, like brain-derived neurotrophic factor (BDNF), nerve growth factor (NGF), or insulin-like growth factor (IGF-1), is increased in the brain after a regimen of daily physical exercise in animals (Carro, Nunez, Busiguina, & Torres-Aleman, 2000; Gobbo & O'Mara, 2005; Neeper, Gomez-Pinilla, Choi, & Cotman, 1995; Neeper, Gomez-Pinilla, Choi, & Cotman, 1996). The amount of neurotrophic factor release correlated with faster learning and better retention over a period of 1 week (Vaynman, Ying, & Gomez-Pinilla, 2004). Exercise also enhances neurogenesis (van Praag et al., 1999; van Praag, Kempermann, & Gage, 1999), which could also contribute to better learning.

In humans, the association between learning improvement and exercise-induced humoral changes has not yet been investigated. It has only been shown that the P300 component of the event-related brain potential (ERP) has a larger amplitude and a shortened latency in attentional challenging tasks, consistent with an overall arousing effect, after a short bout of anaerobic exercise compared to rest (Hillman, Snook, & Jerome, 2003; Magnie et al., 2000; Nakamura, Nishimoto, Akamatu, Takahashi, & Maruyama, 1999). Furthermore, peripheral catecholamine levels may increase after physical exercise (Hyyppa, Aunola, & Kuusela, 1986; Koch, Johansson, & Arvidsson, 1980; Kraemer et al., 1999; Musso, Gianrossi, Pende, Vergassola, & Lotti, 1990), but several other studies found no changes (Bracken, Linnane, & Brooks, 2005; Hartling, Kelbaek, Gjorup, Nielsen, & Trap-Jensen, 1989; Sothmann, Gustafson, & Chandler, 1987). The only study probing central dopamine level changes after a single bout of exercise by positron emission tomography yielded negative results (Wang et al., 2000).

Outside the realm of exercise research, increased peripheral epinephrine levels were correlated with enhanced memory performance in both animals (Costa-Miserachs, Portell-Cortes, Aldavert-Vera, Torras-Garcia, & Morgado-Bernal, 1994) and humans (Cahill & Alkire, 2003). Together, these findings suggest that an exercise-induced increase of catecholamines and neurotrophic factors might improve learning.

We here examined the effects of single bouts of controlled intense anaerobic or moderate aerobic physical exercises (lactate levels above 10 mmol/l or below 2 mmol/l, respectively; Spurway, 1992) on learning and memory. We chose a language learning model because lexical learning is an important aspect of every day life. In search of the mediating mechanisms, we additionally assessed exercise-induced changes in mood, peripheral catecholamine plasma levels and BDNF serum levels and correlated these parameters with subjects' cognitive performance.

## 2. Materials and methods

### 2.1. Subjects

A total of 30 healthy male sport students (mean age: 22.2 ± 1.7 years; range: 19–27) participated as subjects in this prospective randomized controlled trial. Two subjects failed to complete the study due to exercise injuries unrelated to the study. Another subject failed to learn, presumably due to inattentive responding (reactions times on average <300 ms). Therefore, data analysis was conducted with a total of 27 subjects[2]. Participants' written informed consent was obtained according to the declaration of Helsinki. The Ethical Committee of the University of Muenster had approved the physiological intervention of the study. All subjects were native German speakers and right-handed as assessed by the Edinburgh Handedness Inventory (Oldfield, 1971). They had at least 13 years of formal education.

Exclusion criteria comprised bilingualism, a history of neurological, psychiatric or medical diseases, acute infections, intake of medications affecting the central nervous system, recent consumption of recreational drugs as assessed by urinary drug screening, smoking >10 cigarettes/day or drinking >6 cups of coffee/day or >50 g alcohol (equivalent of two glasses of wine) consumption/day.

### 2.2. Study design

#### 2.2.1. Preexamination

*2.2.1.1. Cognitive screening.* Subjects were screened with a comprehensive neuropsychological test battery prior to participation. This battery comprised tests of general intellectual functioning, attention, verbal fluency, digit spans, verbal and visuospatial memory, and personality scales. Additionally, two questionnaires assessed possible exercise dependency (Hausenblas & Downs, 2002) and quantified the average amount of physical

[2] For the retention test 8–10 months after the training, only 25 subjects were available.

Table 1
Means and standard deviations of the neuropsychological background measures and correlation coefficients (Pearson) with training success on the novel vocabulary

| Test | Mean | *SD* | Correlation with learning success (*r*), condition | | |
|---|---|---|---|---|---|
| | | | "relaxed" | "moderate" | "intense" |
| Edinburgh Handedness Inventory (laterality index) | 85.9 | 16.0 | −0.17 | −0.15 | −0.32 |
| Number of languages spoken fluently | 1.6 | 0.6 | −0.20 | −0.14 | −0.06 |
| VLMT: list A leaning success (block 5 minus 1) | 5.4 | 1.4 | −0.13 | −0.29 | −0.19 |
| VLMT: immediate free recall (PR) | 77.0 | 13.8 | 0.17 | 0.18 | 0.10 |
| VLMT: interference list B (PR) | 58.5 | 34.9 | 0.06 | 0.33 | 0.32 |
| VLMT: delayed free recall (PR) | 62.2 | 30.5 | 0.28 | 0.21 | 0.19 |
| WMS verbal paired associates: sum of blocks 1–3 | 22.3 | 1.7 | 0.36 | 0.39[a] | 0.23 |
| WMS verbal paired associates: delayed recall | 7.6 | 0.6 | 0.31 | 0.37 | 0.42[a] |
| WMS visual paired associates: sum of blocks 1–3 | 15.7 | 2.5 | 0.36 | 0.37 | 0.18 |
| WMS visual paired associates: delayed recall | 5.9 | 0.3 | 0.12 | 0.09 | 0.26 |
| Rey-Figure, Copy | 34.9 | 1.0 | 0.10 | 0.10 | −0.04 |
| Rey-Figure, delayed recall | 24.7 | 4.1 | 0.05 | 0.32 | 0.32 |
| RWT: Word fluency (mean PR) | 32.9 | 14.7 | 0.33 | 0.32 | 0.15 |
| Trail Making A (PR) | 54.8 | 25.6 | 0.21 | 0.32 | 0.15 |
| Trail Making B (PR) | 57.0 | 28.0 | 0.05 | 0.03 | −0.25 |
| WAIS-R: Vocabulary (WP) | 9.8 | 1.9 | 0.22 | 0.23 | 0.23 |
| WAIS-R: Similarities (WP) | 11.7 | 2.2 | −0.30 | −0.03 | −0.33 |
| WAIS-R: Picture Completion (WP) | 11.8 | 1.7 | 0.07 | 0.09 | −0.15 |
| WAIS-R: Block design (WP) | 11.2 | 2.9 | 0.22 | 0.44[a] | 0.06 |
| digit span forward (PR) | 66.3 | 23.8 | 0.37 | 0.04 | 0.27 |
| digit span backward (PR) | 59.2 | 30.7 | 0.07 | 0.06 | −0.20 |
| Corsi block tapping forward (PR) | 70.2 | 26.9 | 0.10 | −0.08 | −0.24 |
| Corsi block tapping backward (PR) | 72.7 | 20.1 | 0.25 | 0.13 | −0.01 |
| Sensation Seeking Scale (total score) | 21.6 | 3.4 | −0.16 | −0.31 | −0.28 |
| BDI scores | 3.0 | 3.0 | 0.20 | 0.00 | 0.27 |
| Neo-FFI: neuroticism | 1.3 | 0.5 | 0.14 | −0.08 | 0.09 |
| Neo-FFI: extroversion | 2.5 | 0.4 | 0.23 | 0.10 | 0.17 |
| Neo-FFI: openness to experience | 2.2 | 0.5 | 0.18 | 0.39[a] | 0.50[b] |
| Neo-FFI: agreeableness | 2.6 | 0.4 | −0.44[a] | −0.24 | −0.06 |
| Neo-FFI: conscientiousness | 2.6 | 0.5 | 0.23 | 0.49[a] | 0.31 |
| STAI Trait (PR) | 52.3 | 19.8 | 0.41[a] | 0.18 | 0.34 |
| FfkA (total activity) | 17.3 | 11.0 | 0.19 | 0.21 | 0.26 |
| EDS (total score) | 2.9 | 0.6 | 0.19 | 0.32 | −0.08 |

VLMT = Verbaler Lern- und Merkfaehigkeitstest (German version of the Rey Auditory Verbal Learning Test); RWT = Regensburger Wortfluessigkeitstest (German version of the Controlled Oral Word Association Test); WMS = Wechsler Memory Scale (German version); WAIS = Wechsler Adult Intelligence Scale (German version); Neo-FFI = German version of the Neo Five Factor Inventory; FFkA = "Freiburger Fragebogen zur koerperlichen Aktivitaet" (German questionnaire of the average amount of physical activity per week; Frey et al., 1999); EDS = Exercise Dependence Scale (Hausenblas & Downs, 2002); *SD* = standard deviation; PR = percent rank; WP = Wechsler points.

[a] Significant correlation on a 0.05-level (without correction for multiple testing).

[b] Significant correlation on a 0.01-level (without correction for multiple testing).

activity per week (Frey, Berg, Grathwohl, & Keul, 1999). Cognitive measures were all within the normal range. The results of the neuropsychological tests are shown in Table 1.

*2.2.1.2. Fitness test.* Physical fitness levels were determined by a field test on a 200 m track. The exercise test started at a speed of 8 km/h. Every 3 min, running speed was enhanced by 2 km/h until exhaustion (generally at 18 km/h). On each speed level, participants received continuous acoustic signals in a given frequency as pacing signals. For the determination of lactate concentrations, capillary blood samples were taken from the ear lobe immediately after each speed level as well as after 3 and 6 min during the recovery phase afterwards. At the same time samples, subjective ratings of the perceived exertion were obtained using the Borg-Scale (Borg 1975 in Nybo, Nielsen, Blomstrand, Moller, & Secher, 2003), ranging from 6 (no exhaustion at all) to 20 (complete exhaustion).

*2.2.2. Exercise interventions*

Using a cross-over design, every subject took part in three conditions on different days, spaced at least 1 week apart (see Fig. 1). The conditions differed with regard to the intensity of physical activity. The condition "relaxed" served as a control and consisted of 15 min being sedentary. The condition "moderate" consisted of 40 min of low impact running at a fixed individual heart rate. The individual target heart rate was based on the results of the initial physical fitness test and ensured that lactate levels remained below 2 mmol/l (aerobic condition). In the condition "intense", subjects performed two sprints of 3 min each, separated by a 2 min break. Each sprint started at 8 km/h, increased every 10 s by 2 km/h, until exhaustion. This was an anaerobic condition with lactate levels greater than 10 mmol/l.

The sequence of the three conditions was randomized across subjects. For the moderate and intense conditions, subjects also rated their perceived exertion on the Borg-Scale immediately after the intervention. Heart rates were assessed prior to and after each of the interventions (2 time samples per condition). Vocabulary learning started 15 min after the respective intervention.

Peripheral levels of BDNF and catecholamines (dopamine, epinephrine, and norepinephrine), lactate levels, and mood ratings were assessed prior to as well as following each of the interventions and immediately after vocabulary learning. For mood ratings, the Positive and Negative Affective Schedule (PANAS) was used (Watson, Clark, & Tellegen, 1988; German version: Krohne, Egloff, Kohlmann, & Tausch, 1996), comprising

ten positive and ten negative adjectives, which measure the dimensions positive affect (high score: a state of high energy, low score: sadness and lethargy) and negative affect (high score: state of distress; low score: state of calmness).

#### 2.2.3. Language learning paradigm

The detailed structure of the vocabulary learning task has been described elsewhere (Breitenstein et al., 2005; Breitenstein, Kamping, Jansen, Schomacher, & Knecht, 2004; Breitenstein & Knecht, 2002; Knecht et al., 2004). The learning principle was associative learning: the "correct" pairings of an visually presented daily object and a novel word (e.g., car and /glump/) co-occur over the course of the five training blocks ten times more often as "incorrect" pairings (e.g., bike and /glump/), which are shown only once (Breitenstein & Knecht, 2002). For each subject and each condition, there were a total of 600 training trials (5 blocks × 120 trials). Each trial consisted of a visually presented object picture, presented 200 ms after the onset of the auditory presentation of a novel word (pseudoword, all normalized to a duration of 600 ms). During picture presentation, which lasted for 1 s, subjects had to press one of two keys with their right hand on a response pad to indicate whether the pairing was correct or not. To prevent subjects from reflecting on their responses, the intertrial interval was limited to 1 s. The instruction was to "intuitively decide if objects and novel words match or not".

Subjects were told that only responses occurring in the 1 s interval of picture presentation were accepted for data analysis. They were not informed about the underlying frequency principle.

Subjects' ability to translate the novel words into German was tested in a transfer test immediately after the training session. During this transfer test (1 block with 120 trials), German object names were acoustically presented in pairs with one of the spoken pseudowords. Subjects had to decide whether the pairing was correct or not. The transfer test was administered again 1 week and >8 months after the last training day to assess retention of the vocabulary. On the 1 week retention session, subjects had to also name each picture in the novel vocabulary by writing down the correct novel word (free recall test).

A different version of the novel vocabulary was used for every condition. The three sets were matched for frequency, number of syllables, and familiarity of the German objects names and for number of syllables, associations with existing words, and acoustic valence of the pseudowords.

Dependent variable were learning speed (increase of correct responses from block 1 to block 5) and the overall learning success (performance on the transfer and the free recall tasks) immediately after the training, at the retention sessions 1 week and 6–8 months post, respectively. In addition to accuracy, response times during the vocabulary training were analyzed.

#### 2.2.4. Biochemical blood analyses

Lactate concentration in capillary blood was measured by a photometric method using a commercially available kit (EKF Diagnostic, Magdeburg, Germany).

The HPLC assay kit for plasma catecholamines was provided by Chromsystems (Munich, Germany). Reversed-phase chromatography was performed on an isocratic Kontron 422 liquid chromatograph (Neufahrn, Germany), interfaced with a model 41,000 electrochemical detector (Chromsystems, Munich, Germany). Complete blood cell count, including an automated differential, was performed by a SE9500 automated full blood count analyzer (Sysmex Deutschland GmbH, Norderstedt, Germany). BDNF level in serum was measured using an ELISA kit (Quantikine human BDNF, R&D Systems, Wiesbaden, Germany).

Dopamine and epinephrine blood plasma levels below the biochemical detection threshold (30 ng/l for dopamine, 15 ng/l for epinephrine) were set to a mean value between zero and the respective detection threshold (15 ng/l for dopamine, 7.5 ng/l for epinephrine).

All BDNF and catecholamine blood plasma levels were corrected for changes in overall blood volume (Dill & Costill, 1974).

### 2.3. Data analysis

Behavioral (accuracy and reaction times on the vocabulary learning task) and biochemical blood data and mood ratings were analyzed using repeated measures ANOVAs, with polynomial contrasts on the factor training block (blocks 1–5) or time sample (3: prior to exercise, after exercise, after vocabulary training). Greenhouse–Geisser adjustments were applied in analyses with two or more degrees of freedom. Significant interactions or main effects were followed up by paired $t$-tests in case of significance. For clarity of presentation, only effects involving the factor exercise condition (relaxed, moderate, intense) are presented.

Immediate and delayed (1 week and >8 months post) retention data of the vocabulary were analyzed separately using univariate ANOVAs with the repeated factor exercise condition.

Correlations were analyzed using Pearson's correlations coefficients. Because of the explorative nature of the present study, it was considered appropriate to omit Bonferroni-corrections of significance levels.

## 3. Results

### 3.1. Learning performance

Using a cross-over study design, learning performance was assessed directly after high impact anaerobic sprints (intense condition), low impact aerobic running (moderate condition), or a period of rest (relaxed condition).

#### 3.1.1. Accuracy

There was a difference in learning speed between these three conditions (condition × block: quadratic trend, $F(1,26) = 9.39$, $p = .005$). Post hoc tests showed that learning speed across the five training blocks was significantly faster after intense running compared to being sedentary (condition × block: quadratic trend, $F(1,26) = 9.39$, $p = .005$) and moderate running (condition × block: quadratic trend, $F(1,26) = 5.27$, $p = .03$) conditions (see Fig. 2). Please note that performance in the intense condition was already at ceiling in block 4. For the other two conditions, subjects needed one more training block (block 5) to reach a comparable level of performance. This indicates that learning was accelerated by 20 percent in the intense condition.

The transfer sessions immediately after training, 1 week and >8 months post training, respectively, were not significantly different for the three conditions (main effects of condition: all $p > .18$). However, exploratory comparisons showed that subjects presented with a better 1-week retention for the intense as compared to the moderate condition ($t(26) = 2.24$, $p = .03$; see Fig. 2). The free recall naming task at the 1-week post assessment did not yield significant differences between the three conditions (main effect of condition: $p = .57$).

#### 3.1.2. Analysis of training reaction times

Only the main effect of condition yielded significance ($F(2,52) = 5.11$, $p = .01$). Post hoc tests revealed significantly faster responding in the condition "intense" compared to both "relaxed" and "moderate" (both $F(1,26) > 4.35$, $p < .05$; see Fig. 3). However, overall learning success (accuracy on block 5 minus block 1) was neither correlated with the mean reaction time on block 1 nor with the mean

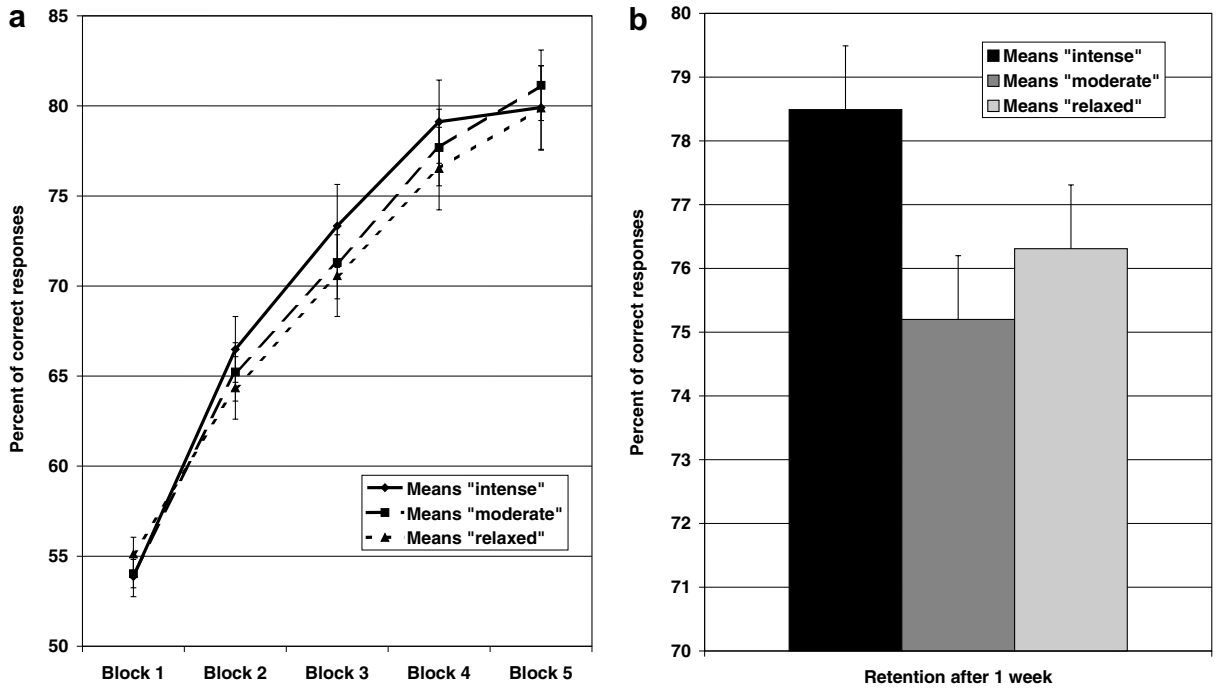


Fig. 2. (a) Vocabulary learning success (means ± SEM) across the five training blocks was faster after intense physical exercise (solid line) compared to after moderate exercise (dashed line) or the relaxed condition (dotted line). (b) Retention (means ± SEM) of the vocabulary after 1 week was better in the "intense" condition (black) as compared to the "moderate" condition (dark gray).

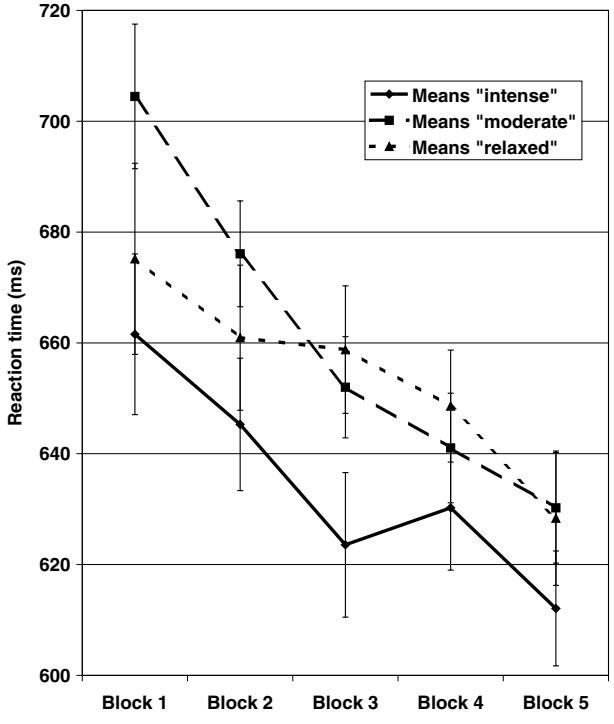


Fig. 3. Mean reaction times (±SEM) in ms across the five training blocks. Responses were faster after intense exercise (solid line) compared to both moderate exercise (dashed line) and being sedentary (dotted line).

reaction time across all blocks, indicating that unspecific motor arousal after intense physical exercise cannot explain the superior learning rates.

### 3.2. Measures of unspecific arousal

Measures of unspecific arousal were assessed to control for unspecific effects mediating the exercise-induced learning improvement.

#### 3.2.1. Heart rate

In the moderate condition, subjects' heart rates were in the range of 110 and 160 beats/min (median: 140 beats/min). After intense exercise the heart rates were in the range of 163 and 202 beats/min (median: 184 beats/min).

There was a significant interaction of time sample (baseline, post intervention) and condition (linear trend: $F(1,26) = 1680.63$, $p < .001$). Post hoc analyses showed that heart rates increased post intervention only for the moderate (mean increase of 67.3, *SD*: 14.3 beats/min) and intense (mean increase of 111.1, *SD*: 10.8 beats/min) conditions (both $t(26) > |24.52|$, $p < .001$), but not for the relaxed condition. Baseline heart rates did not differ between the conditions. There were no correlations between exercise-induced heart rate changes and learning outcome, although baseline heart rates prior to intense exercise were correlated with immediate and delayed (1 week/>8 months) retention outcomes (all $r > .41$, $p < .03$).

#### 3.2.2. Blood lactate concentrations

The analyses of the peripheral lactate levels revealed a significant interaction of condition and time of measurement (quadratic trend: $F(1,26) = 1002.42$, $p < .001$). As

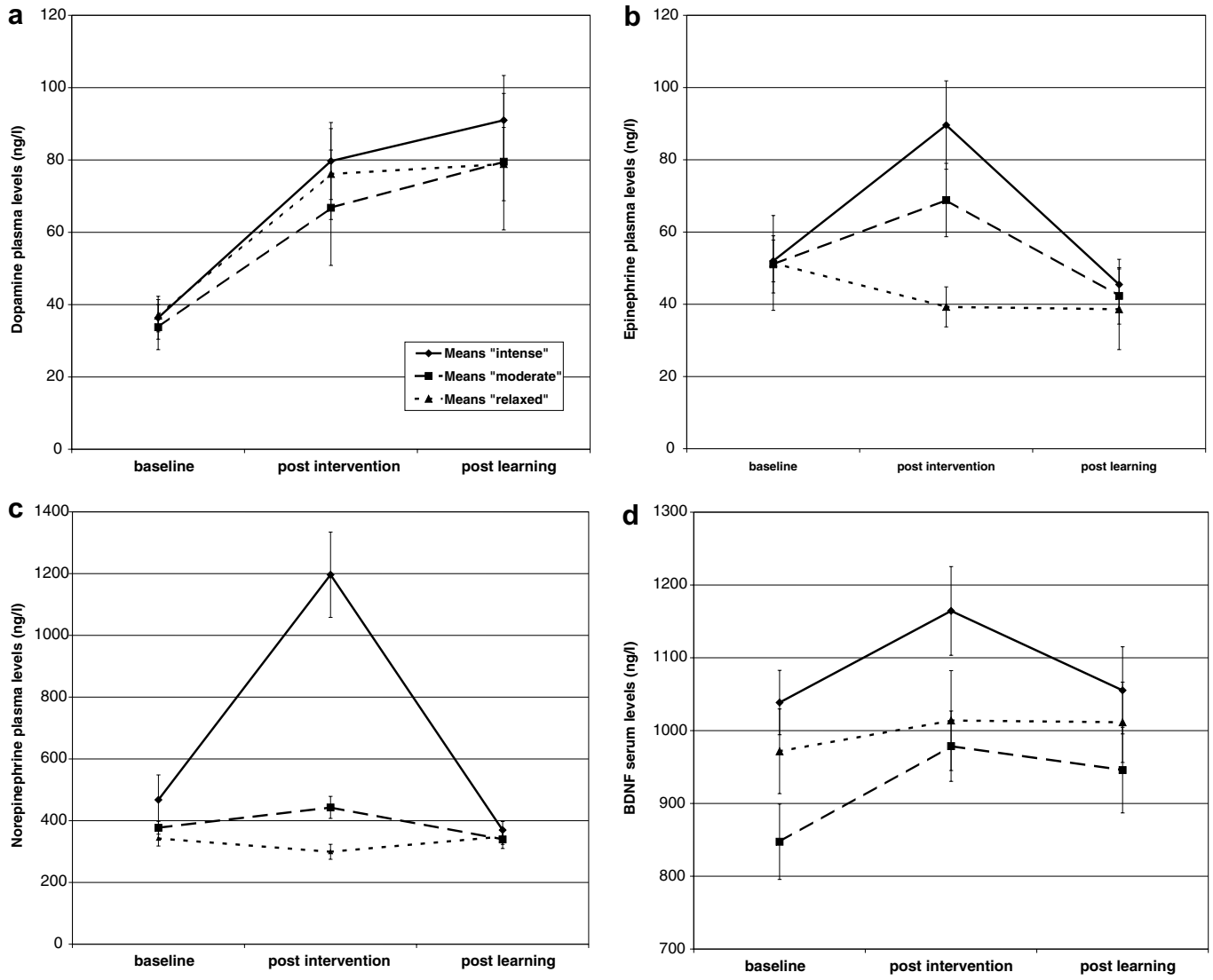


Fig. 4. Exercise-induced changes in blood plasma levels (±SEM, corrected for differences in overall blood volume) of dopamine (a), epinephrine (b), norepinephrine (c) and blood serum levels of BDNF (d) for the conditions "relaxed" (dotted line), "moderate" (dashed line) and "intense" (solid line) across the three time samples (baseline, post intervention, post learning).

expected, lactate levels showed a greater exercise-induced increase during the intense as compared to each of the other two conditions (interaction of condition × time sample: quadratic trend, both $F(1,26) > 1002{,}42$, $p < .001$). There were also differences at baseline ($F(2,52) = 4.62$, $p = .02$; "moderate" > "relaxed, $t(26) = 2.84$, $p = .009$). There were no correlations between exercise-induced lactate level changes and learning outcome.

### 3.3. Exercise-induced neurotransmitter changes: Catecholamine blood plasma levels

Peripheral catecholamine plasma levels (dopamine, epinephrine, norepinephrine) were assessed at baseline, immediately after the respective intervention, and after vocabulary learning to determine possible contributions to the boost in learning.

#### 3.3.1. Dopamine

There were significant linear increases in dopamine plasma levels across the three time samples for all three conditions (main effect of time sample: linear trend, $F(1,26) = 30.66$, $p < .001$). Visual inspection of the data suggested that plasma dopamine levels showed a steeper increase in the intense as compared to the other two condition, but the interaction of condition and time sample did not yield significance (see Fig. 4a). Baseline dopamine plasma levels were not different for the three conditions.[3]

[3] Two subjects were excluded because their dopamine levels were greater than two times the *SD* of the group mean ($n = 25$). Values below the detection threshold were substituted (see Section 2). Separate analyses for the subsample of subjects with values above the biochemical detection threshold ($n = 11$) also yielded a main effect of time sample (linear trend, $F(1,10) = 15.77$, $p = .003$) and no interaction of condition and time sample.

*3.3.2. Epinephrine*

The ANOVA yielded a significant interaction of condition and time sample (quadratic trend, $F(1,24) = 7.03$, $p = .01$). Post hoc tests showed stronger exercise-induced changes for the intense as compared to the relaxed condition (condition × time sample: quadratic trend, $F(1,24) = 7.03$, $p = .01$). Additional post hoc analyses for each level of time sample separately yielded a main effect of condition only for the time sample after the intervention ($F(2,48) = 6.67$, $p = .006$; "moderate", "intense" > "relaxed", both $t(24) > |3.01|$, $p < .006$). There were no baseline differences between conditions (see Fig. 4b).[4]

*3.3.3. Norepinephrine*

There was a significant interaction between condition and time sample (quadratic trend: $F(1,26) = 31.32$, $p < .001$). Post hoc analyses showed that the intense condition yielded significantly steeper norepinephrine changes after exercise as compared to the other two conditions (interaction of condition × time sample: quadratic trend, both $F(1,26) > 27.79$, $p < .001$). Furthermore, moderate exercise led to steeper increases as compared to being sedentary (interaction of condition × time sample: quadratic trend, $F(1,26) = 12.81$, $p = .001$). Separate analyses for each time sample separately showed that the conditions differed only at the time sample immediately post intervention ($F(2,52) = 39.46$, $p < .001$; "intense" > "moderate" > "relaxed", all $t(26) > |3.51|$, $p \leqslant .002$). There were no significant differences at baseline (see Fig. 4c).

*3.4. Exercise-induced changes in BDNF blood serum levels*

To also determine the possible contribution of neurotrophic factors in exercise-induced learning enhancement, peripheral BDNF serum levels were determined at baseline, immediately after the respective exercise, and after vocabulary training. BDNF baseline values differed for the three conditions ($F(2,52) = 5.88$, $p = .006$; intense > moderate: $t(26) = -3.33$, $p = .003$). We will therefore only interpret significant slope differences across the three time samples between conditions. The ANOVA yielded a significant interaction of condition and time sample (quadratic trend, $F(1,26) = 4.38$, $p = .05$). Post hoc tests revealed significantly stronger changes across times samples for the intense as compared to the relaxed condition (interaction of condition × time sample: quadratic trend, $F(1,26) = 4.38$, $p = .05$; see Fig. 4d).

*3.5. Correlations between behavioral and physiological parameters*

To examine the association of learning success with exercise-induced changes in physiological parameters, we calculated the correlations of learning indicators (overall learning success and retention outcome after 1 week and after 8–10 months) with the respective changes in physiological parameters. There were no significant correlations with norepinephrine blood plasma level changes. For the other physiological parameters, the following significant correlations emerged:

A decrease in dopamine blood plasma levels during learning (post intervention minus post learning) was related to a better retention outcome immediately after learning and after 1 week in the intense condition (both $r > .46$, $p < .02$, see Figs. 6a and b)[5]. The absolute dopamine concentration after intense exercise also predicted the immediate and 1 week delayed retention outcomes (both $r > .33$, $p < .10$)[5]. Because subjects with higher absolute dopamine concentrations prior to learning (post exercise) also showed the strongest dopamine level decreases during learning ($r = .85$, $p < .001$), absolute dopamine concentrations during learning may be the crucial factor for the enhanced immediate/intermediate learning outcome.

For epinephrine exercise-induced blood plasma changes, there were no significant correlations with learning outcome. Because absolute concentrations of epinephrine could be more important for memory consolidation than relative increases from baseline, we correlated the absolute epinephrine blood plasma concentrations after exercise with immediate and delayed learning success. Only the epinephrine concentrations after the intense intervention (prior to learning) correlated with long-term retention of the vocabulary (retention after >8 months $r = .41$, $p < .05$, see Fig. 6c)[6]. A trend was also seen for the correlation between epinephrine concentrations post intense exercise and vocabulary retention after 1 week ($r = .38$, $p = .06$)[6].

To additionally support the link between epinephrine changes and learning success, we contrasted subjects with low versus high exercise-induced epinephrine changes in the intense condition (median split using the peripheral epinephrine blood plasma levels after intense exercise). These analyses yielded a significantly better outcome for the "high" as compared to the "low" epinephrine changes group for the retention results after 1 week and after >8 months ($t(23) = -2.60$, $p = .02$ and $t(22) = -2.27$, $p = .03$, respectively, see Fig. 5)[6].

For BDNF serum levels, the more sustained the BDNF levels during learning (BDNF levels after learning minus

[4] Two subjects were excluded because their epinephrine levels were greater than two times the *SD* of the group mean ($n = 25$). Values below the detection threshold were substituted (see Section 2). Separate analyses for the subsample of subjects with values above the biochemical detection level ($n = 11$) yielded a significant interaction of condition and time sample (quadratic trend, $F(1,10) = 7.75$, $p = .02$). Post hoc tests only showed a difference between the conditions "relaxed" and "intense" (condition × time sample: quadratic trend, $F(1,10) = 7.75$, $p = .02$), but not between the other pairwise comparisons. Additional post hoc analyses for each level of time sample separately yielded a main effect of condition only for the time sample after the intervention ($F(2,20) = 4.70$, $p = .05$; "moderate", "intense" > "relaxed", both $t(10) > |3.03|$, $p < .01$).

[5] Two subjects were excluded because their dopamine levels were greater than two times the *SD* of the group mean ($n = 25$).

[6] Two subjects were excluded because their epinephrine levels were greater than two times the *SD* of the group mean ($n = 25$ or 24 for the retention after >8 months, see Section 2).

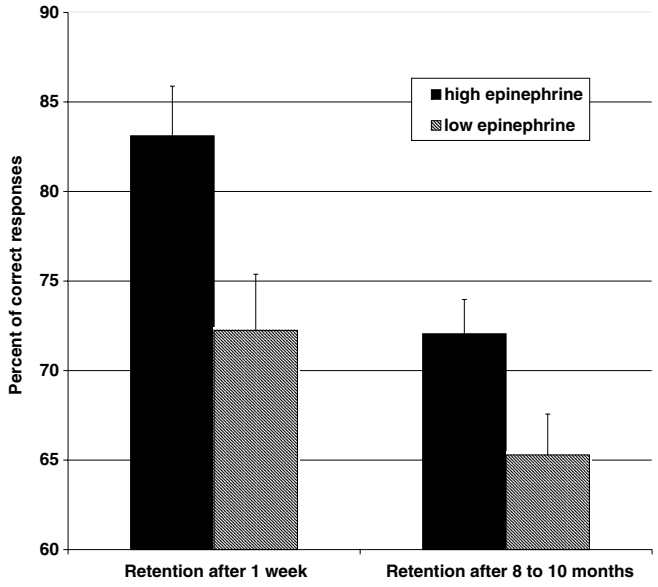


Fig. 5. Differences in retention outcome (means ± SEM) after 1 week and after 8–10 months for subjects with high and low epinephrine blood plasma levels after intense exercise (split-half method).

BDNF levels after exercise), the greater the immediate learning success (accuracy on block 5 minus block 1) for the "intense" condition ($r = .38$, $p = .05$; see Fig. 6d).

### 3.6. Analyses of positive and negative mood ratings

There were no significant interactions or main effects involving the factor condition for negative mood ratings.

The analysis of the positive mood ratings yielded a significant interaction of condition and time sample (quadratic trend, $F(1,25) = 11{,}26$, $p = .003$[7]). Ratings at baseline and after learning did not differ significantly between the conditions. Ratings differed only for the time sample after the respective interventions ($F(2,50) = 6.52$, $p = .004$), with significantly higher positive mood ratings for the condition "intense" versus "relaxed" as well as for "moderate" versus "relaxed" (both $t(25) > |2.94|$, $p \leqslant .007$[8]).

Furthermore, in the intense conditions, a more sustained exercise-induced increase in positive mood (mood rating post intervention minus mood rating immediate after learning) was marginally associated with an overall better learning success (accuracy on block 5 minus block 1; $r = -.37$, $p = .06$), a better outcome at the immediate transfer session ($r = -.33$, $p = .10$) and with a better retention after 1 week ($r = -.37$, $p = .06$). Changes in mood ratings were not correlated with the changes in peripheral epinephrine plasma levels, indicating that both factors contributed independently to better learning.

[7] $n = 26$ because one subject did not entirely complete one of the PANAS questionnaires.

[8] Two subjects were excluded because their epinephrine levels were greater than two times the *SD* of the group mean ($n = 25$ or 24 for the retention after >8 months, see Section 2).

## 4. Discussion

The main finding of the present study was that intense exercise directly improves learning: After two sprints of less than 3 min each, subjects learned 20 percent faster compared to moderate exercise or being sedentary. To our knowledge, this is the first study of immediate exercise-induced effects on a complex learning task with a parallel analysis of neurophysiological correlates (changes in peripheral catecholamine or BDNF levels) in humans. Our results suggests that short bouts of exercise could be used in situations which require an immediate boost of learning, e.g., immediately prior to study phases in children with and without learning deficits.

We were further able to elucidate at least some of the underlying neurophysiological mechanisms of improved learning through exercise. Intense running led to elevated levels of peripheral catecholamines (dopamine, epinephrine, norepinephrine) and BDNF. More sustained BDNF levels during learning (levels after intense exercise minus levels after learning) were related to better short-term learning success, and absolute dopamine and epinephrine levels after intense exercise were related to better intermediate (dopamine) and long-term (epinephrine) retentions of the novel vocabulary. The latter finding was endorsed by the observation, that subjects with relatively higher (as compared to the group mean) epinephrine blood plasma levels after intense exercise had a better long-term retention of the trained vocabulary up to >8 months. We will discuss these findings in more detail below.

### 4.1. Short-term learning improvement and BDNF

We found that more sustained BDNF blood serum levels during learning predicted immediate learning success after intense exercise. This finding is consistent with prior work showing that BDNF secretion is increased after exercise in animals (Neeper et al., 1995; Vaynman et al., 2004) and in humans (Gold et al., 2003). BDNF plays an import role in learning due to its involvement in long-term potentiation in the hippocampus (Pang et al., 2004). BDNF also contributes to synaptic efficacy, neuronal connectivity, and brain plasticity (McAllister, Katz, & Lo, 1999; Schinder & Poo, 2000; Vaynman et al., 2004). The release of other neurotrophic factors like IGF-1 is increased after physical exercise in humans as well (Carro et al., 2000). We chose BDNF because it has been shown to be elevated after physical exercise in humans (Gold et al., 2003), and most of the genes up-regulated after exercise and relevant to plasticity are associated with BDNF (Molteni, Ying, & Gomez-Pinilla, 2002). We could only assess peripheral BDNF levels, but there seems to be an influx of natural BDNF from the blood into the brain (Pan, Banks, Fasold, Bluth, & Kastin, 1998); although, this is challenged by other authors (Sakane & Pardridge, 1997; Wu, 2005). Our findings show that BDNF could be one mediator between physical

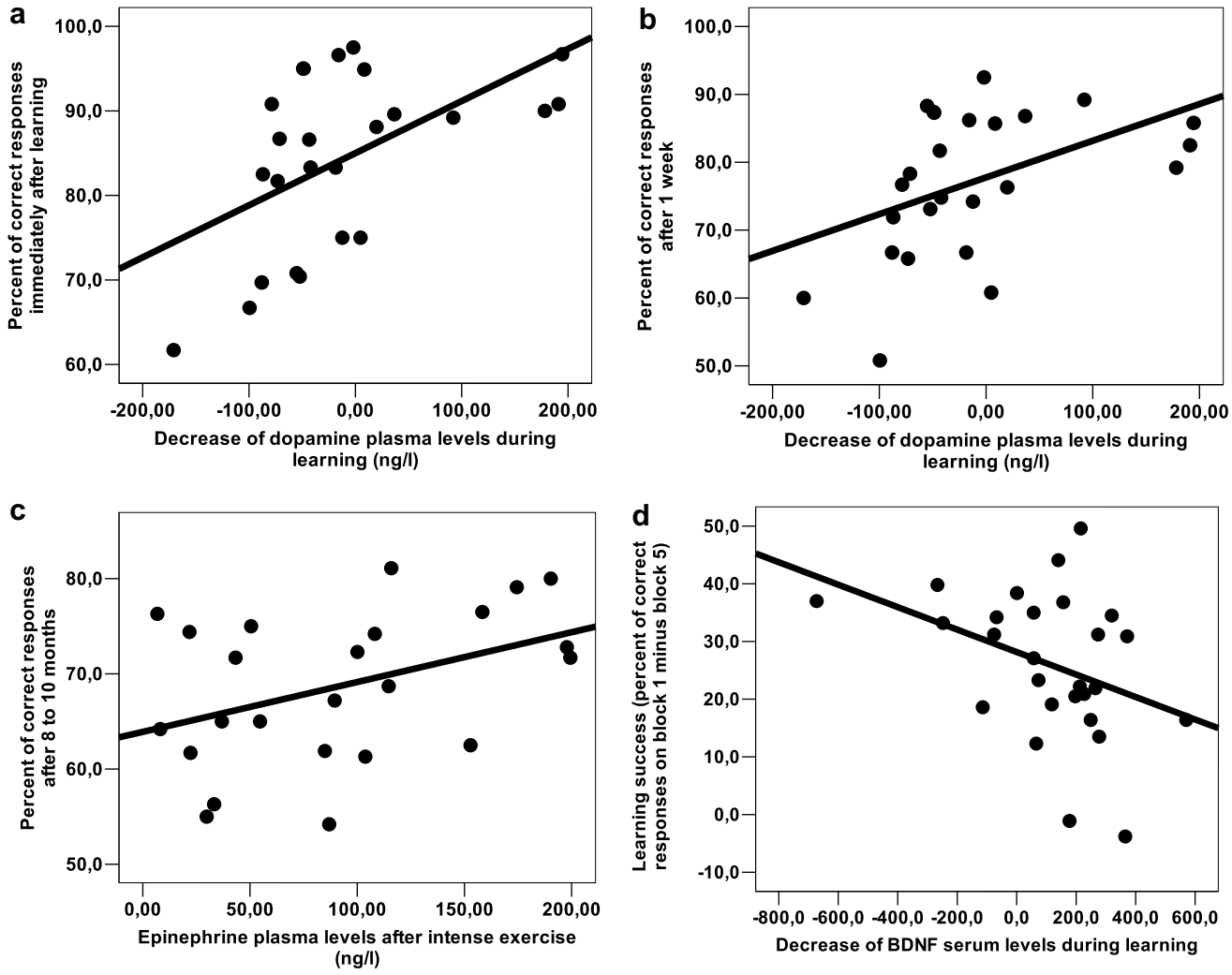


Fig. 6. (a,b) Peripheral dopamine plasma levels during learning (after intense intervention minus after learning) predicted immediate retention outcome (a) and retention outcome after 1 week (b). (c) Peripheral epinephrine plasma levels post intense exercise correlated with the retention outcome after >eight months; (d) BDNF serum levels during learning (levels after intense intervention minus levels after learning) were related to immediate learning success (accuracy on block 5 minus block 1).

exercise and learning improvement (Cotman & Berchtold, 2002; Vaynman et al., 2004).

### 4.2. Intermediate learning improvement and dopamine

We observed a correlation between greater decreases of dopamine blood plasma levels during learning and better intermediate retention after intense exercise. This may seem counterintuitive at first sight, but further analyses showed that higher *absolute* dopamine levels in the initial stages of learning drove the correlation of an enhanced retention outcome.

The role of dopamine for learning has been variously demonstrated (Fiorillo, Tobler, & Schultz, 2003; Floel et al., 2005; Jay, 2003; Knecht et al., 2004; Wise, 2004). Dopamine is part of the internal reward system (Schultz, 2002), regulates the prefrontal cortical circuitry underlying working memory (Castner & Goldman-Rakic, 2004; Marie & Defer, 2003) and modulates arousal and attention (Nutt & Fellman, 1984). It even seems to be critical for neurogenesis (Baker, Baker, & Hagg, 2004).

Animal studies suggest an increase in dopamine release in the brain during exercise (Hattori et al., 1994; Meeusen & De Meirleir, 1995; Sutoo & Akiyama, 2003). Our controlled intense exercise condition also led to elevated levels of peripheral dopamine levels, contrary to the results of other studies in humans with less objective criteria for physical exertion levels (Bracken et al., 2005; Kraemer et al., 1999). Peripheral dopamine levels are less than perfect correlates of dopamine release in the brain, because dopamine cannot cross the blood-brain barrier. We had to operate on the assumption that brain and systemic dopamine levels responded similarly to physical exercise. Thus, an enhanced presynaptic availability of endogenous dopamine could have led to stronger phasic dopamine signals coding for stimulus salience (Breitenstein, Korsukewitz K, Floel A, Kretzschmar T, & Diederich K, 2006; Schultz, 2002), contributing to better memory consolidation.

### 4.3. Long-term learning improvement and epinephrine

We found a correlation between absolute blood plasma concentrations of epinephrine after intense exercise and long-term retention outcome. There are several explanations for the beneficial influence of exercise on cognitive functions mediated by epinephrine. One is that the increased memory consolidation after intense exercise was driven by increased arousal during learning (Hollmann & Struder, 2000; Sharot & Phelps, 2004) and improved attention (Hillman et al., 2003; Magnie et al., 2000; Nakamura et al., 1999). Heart rates and lactate levels were higher and the reaction times during learning were shorter after the intense exercise as compared to the other two conditions, but neither of these parameters was correlated with learning success. These findings speak against increased arousal as the exclusive mediator of improved learning. Nevertheless, arousal may have mediated part of the effect, because prior studies already demonstrated that absolute peripheral epinephrine during learning contributes to memory consolidation in animals (Costa-Miserachs et al., 1994; Liang, Chen, & Huang, 1995) and in humans (Cahill & Alkire, 2003), with the crucial factor being the grade of arousal during encoding (Cahill & Alkire, 2003).

Peripheral epinephrine does not cross the blood-brain barrier (Bradbury, 1993), but there might be an indirect pathway to influence the central nervous system: One possible route is the activation of vagal afferent fibers via β-adrenergic receptors. This vagal stimulation by peripheral epinephrine leads to an increased neural firing in the noradrenergic connections between the nucleus solitarius and the amygdala (Clayton & Williams, 2000; Miyashita & Williams, 2006) or the hippocampus (Miyashita & Williams, 2004). Enhanced noradrenergic release in this area might then lead to increased general brain excitability with improved learning capability (Boyeson & Feeney, 1990; Feeney & Hovda, 1985; Goldstein, 1999). This indirect pathway could be the neurophysiological link for the observed correlation between peripheral epinephrine plasma levels and enhanced vocabulary retention in the present study.

### 4.4. Aerobic versus anaerobic physical exercise

We did not observe an effect of a single bout of moderate (aerobic) physical exercise on learning and memory. The two exercise interventions differed not only in terms of intensity, but also in duration. However, the intense condition led to a facilitation of learning, despite a *shorter* duration compared to the moderate condition. It also seems unlikely that the lack of learning improvement in the moderate condition was due to the greater fatigue or the longer duration (40 min), because the moderate intervention did not differ from the resting condition in terms of learning success. But we cannot rule out that this was due to an interaction of facilitation and debilitation (Tomporowski & Ellis, 1986). However, after prolonged moderate exercise, there may be effects on mental functioning (e.g., Colcombe et al., 2004). It remains to be determined in future studies whether short high impact and prolonged low impact exercise have comparable effects and are mediated by similar mechanisms. Please note that the three conditions of our study yielded comparable immediate learning outcomes (block 5 of the training), presumably due to ceiling effects of ten subjects in the intense learning condition at block 4. Seven of these subjects subsequently showed a small dip in performance in block 5, probably due to a lack of continuous motivation. It is feasible that a different learning paradigm, which allows for greater differentiation of learning success, would also yield qualitative differences between the conditions.

Our study was designed as a "proof of principle" experiment, probing the effects of a single bout of physical exercise on learning and memory. Because of its pilot character, the sample size was relatively small, and the moderate statistical power may explain why some of the reported effects were only marginally significant. Nevertheless, we could show that exercise accelerates learning and improves long-term retention of the learned material (at least in subjects with high exercise-induced epinephrine levels). This is pertinent to the organization of learning-supportive environments, e.g., in schools (intense exercise during the breaks) and as a possible treatment for cognitive impairments in cardiovascularly stable people. Our results may also be of significance for the development of treatment options for learning-impaired neurological patients (stroke, dementia) because we could determine BDNF, dopamine, and epinephrine as important mediators of the exercise-induced learning improvement. Further investigations are necessary to determine if the observed effects generalize to other learning modalities, like visual-spatial learning.

## Acknowledgments

This work was supported by the Cusanuswerk, the NRW-Nachwuchsgruppe Kn2000 of the Nordrhein-Westfalen Ministry of Education and Research (Foe.1KS9604/0), the Interdisciplinary Center of Clinical Research Muenster (IZKF Projects FG2 and Kne3/074/04), the Volkswagen Stiftung (Az.: I/80 708), as well as the German Ministry of Education and Research (BMBF: 01GW0520).

## References

Abbott, R. D., White, L. R., Ross, G. W., Masaki, K. H., Curb, J. D., & Petrovitch, H. (2004). Walking and dementia in physically capable elderly men. *JAMA: The Journal of the American Medical Association, 292*, 1447–1453.

Allison, D. B., Faith, M. S., & Franklin, R. D. (1995). Antecedent exercise in the treatment of disruptive behavior: a meta-analytic review. *Clinical Psychology: Science & Practice, 2*, 279–304.

Anderson, B. J., Rapp, D. N., Baek, D. H., McCloskey, D. P., Coburn-Litvak, P. S., & Robinson, J. K. (2000). Exercise influences spatial learning in the radial arm maze. *Physiology & Behavior, 70*, 425–429.

Baker, S. A., Baker, K. A., & Hagg, T. (2004). Dopaminergic nigrostriatal projections regulate neural precursor proliferation in the adult mouse subventricular zone. *European Journal of Neuroscience, 20*, 575–579.

Baruch, D. E., Swain, R. A., & Helmstetter, F. J. (2004). Effects of exercise on Pavlovian fear conditioning. *Behavioral Neuroscience, 118*, 1123–1127.

Boyeson, M. G., & Feeney, D. M. (1990). Intraventricular norepinephrine facilitates motor recovery following sensorimotor cortex injury. *Pharmacology Biochemistry and Behavior, 35*, 497–501.

Bracken, R. M., Linnane, D. M., & Brooks, S. (2005). Alkalosis and the plasma catecholamine response to high-intensity exercise in man. *Medicine and Science in Sports and Exercise, 37*, 227–233.

Bradbury, M. W. (1993). The blood–brain barrier. *Experimental Physiology, 78*, 453–472.

Braszko, J. J., Kaminski, K. A., Hryszko, T., Jedynak, W., & Brzosko, S. (2001). Diverse effects of prolonged physical training on learning of the delayed non-matching to sample by rats. *Neuroscience Research, 39*, 79–84.

Breitenstein, C., Korsukewitz, C., Floel, A., Kretzschmar, T., Diederich, K., & Knecht, S. (2006). Tonic dopaminergic stimulation impairs associative learning in healthy subjects. *Neuropsychopharmacology, 31*, 2552–2564.

Breitenstein, C., Jansen, A., Deppe, M., Foerster, A. F., Sommer, J., Wolbers, T., et al. (2005). Hippocampus activity differentiates good from poor learners of a novel lexicon. *Neuroimage, 25*, 958–968.

Breitenstein, C., Kamping, S., Jansen, A., Schomacher, M., & Knecht, S. (2004). Word learning can be achieved without feedback: implications for aphasia therapy. *Restorative Neurology and Neuroscience, 22*, 445–458.

Breitenstein, C., & Knecht, S. (2002). Development and validation of a language learning model for behavioral and functional-imaging studies. *Journal of Neuroscience Methods, 114*, 173–179.

Burghardt, P. R., Fulk, L. J., Hand, G. A., & Wilson, M. A. (2004). The effects of chronic treadmill and wheel running on behavior in rats. *Brain Research, 1019*, 84–96.

Cahill, L., & Alkire, M. T. (2003). Epinephrine enhancement of human memory consolidation: interaction with arousal at encoding. *Neurobiology of Learning and Memory, 79*, 194–198.

Carro, E., Nunez, A., Busiguina, S., & Torres-Aleman, I. (2000). Circulating insulin-like growth factor I mediates effects of exercise on the brain. *Journal of Neuroscience, 20*, 2926–2933.

Castner, S. A., & Goldman-Rakic, P. S. (2004). Enhancement of working memory in aged monkeys by a sensitizing regimen of dopamine D1 receptor stimulation. *Journal of Neuroscience, 24*, 1446–1450.

Cian, C., Barraud, P. A., Melin, B., & Raphel, C. (2001). Effects of fluid ingestion on cognitive function after heat stress or exercise-induced dehydration. *International Journal of Psychophysiology, 42*, 243–251.

Cian, C., Koulmann, N., Barraud, P. A., Raphel, C., Jimenez, C., & Melin, B. (2000). Influences of variations in body hydration on cognitive function: effect of hyperhydration, heat stress, and exercise-induced dehydration. *Journal of Psychophysiology, 14*, 29–36.

Clayton, E. C., & Williams, C. L. (2000). Noradrenergic receptor blockade of the NTS attenuates the mnemonic effects of epinephrine in an appetitive light–dark discrimination learning task. *Neurobiology of Learning and Memory, 74*, 135–145.

Colcombe, S., & Kramer, A. F. (2003). Fitness effects on the cognitive function of older adults: a meta-analytic study. *Psychological Science, 14*, 125–130.

Colcombe, S. J., Kramer, A. F., Erickson, K. I., Scalf, P., McAuley, E., Cohen, N. J., et al. (2004). Cardiovascular fitness, cortical plasticity, and aging. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 3316–3321.

Costa-Miserachs, D., Portell-Cortes, I., Aldavert-Vera, L., Torras-Garcia, M., & Morgado-Bernal, I. (1994). Long-term memory facilitation in rats by posttraining epinephrine. *Behavioral Neuroscience, 108*, 469–474.

Cotman, C. W., & Berchtold, N. C. (2002). Exercise: a behavioral intervention to enhance brain health and plasticity. *Trends in Neuroscience, 25*, 295–301.

Craft, D. H. (1983). Effect of prior exercise on cognitive performance tasks by hyperactive and normal young boys. *Perceptual and Motor Skills, 56*, 979–982.

Davey, C. P. (1973). Physical exertion and mental performance. *Ergonomics, 16*, 595–599.

Dill, D. B., & Costill, D. L. (1974). Calculation of percentage changes in volumes of blood, plasma, and red cells in dehydration. *Journal of Applied Physiology, 37*, 247–248.

Etnier, J. L., Salazar, W., Landers, D. M., Petruzzello, S. J., Han, M., & Nowell, P. M. (1997). The influence of physical fitness and exercise upon cognitive functioning: a meta-analysis. *Journal of Sport & Exercise Psychology, 19*, 249–277.

Feeney, D. M., & Hovda, D. A. (1985). Reinstatement of binocular depth perception by amphetamine and visual experience after visual cortex ablation. *Brain Research, 342*, 352–356.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science, 299*, 1898–1902.

Floel, A., Breitenstein, C., Hummel, F., Celnik, P., Gingert, C., Sawaki, L., et al. (2005). Dopaminergic influences on formation of a motor memory. *Annals of Neurology, 58*, 121–130.

Fordyce, D. E., & Farrar, R. P. (1991). Enhancement of spatial learning in F344 rats by physical activity and related learning-associated alterations in hippocampal and cortical cholinergic functioning. *Behavioural Brain Research, 46*, 123–133.

Frey, I., Berg, A., Grathwohl, D., & Keul, J. (1999). Freiburger Fragebogen zur körperlichen Aktivität - Entwicklung, Prüfung und Anwendung. *Sozial- und Präventivemedizin, 44*, 55–64.

Gobbo, O. L., & O'Mara, S. M. (2005). Exercise, but not environmental enrichment, improves learning after kainic acid-induced hippocampal neurodegeneration in association with an increase in brain-derived neurotrophic factor. *Behavioural Brain Research, 159*, 21–26.

Gold, S. M., Schulz, K. H., Hartmann, S., Mladek, M., Lang, U. E., Hellweg, R., et al. (2003). Basal serum levels and reactivity of nerve growth factor and brain-derived neurotrophic factor to standardized acute exercise in multiple sclerosis and controls. *Journal of Neuroimmunology, 138*, 99–105.

Goldstein, L. B. (1999). Amphetamine-facilitated poststroke recovery [letter; comment]. *Stroke, 30*, 696–698.

Hartling, O. J., Kelbaek, H., Gjorup, T., Nielsen, M. D., & Trap-Jensen, J. (1989). Plasma concentrations of adrenaline, noradrenaline and dopamine during forearm dynamic exercise. *Clinical Physiology, 9*, 399–404.

Hattori, S., Naoi, M., & Nishino, H. (1994). Striatal dopamine turnover during treadmill running in the rat: relation to the speed of running. *Brain Research Bulletin, 35*, 41–49.

Hausenblas, H. A., & Downs, D. S. (2002). How much is too much? The development and validation of the exercise dependence scale. *Psychology & Health, 17*, 387–404.

Hillman, C. H., Snook, E. M., & Jerome, G. J. (2003). Acute cardiovascular exercise and executive control function. *International Journal of Psychophysiology, 48*, 307–314.

Hogervorst, E., Riedel, W., Jeukendrup, A., & Jolles, J. (1996). Cognitive performance after strenuous physical exercise. *Perceptual and Motor Skills, 83*, 479–488.

Hollmann, W., & Struder, H. K. (2000). Brain function, mind, mood, nutrition, and physical exercise. *Nutrition, 16*, 516–519.

Hyyppa, M. T., Aunola, S., & Kuusela, V. (1986). Psychoendocrine responses to bicycle exercise in healthy men in good physical condition. *International Journal of Sports Medicine, 7*, 89–93.

Jay, T. M. (2003). Dopamine: a potential substrate for synaptic plasticity and memory mechanisms. *Progress in Neurobiology, 69*, 375–390.

Kalmijn, S., Foley, D., White, L., Burchfiel, C. M., Curb, J. D., Petrovitch, H., et al. (2000). Metabolic cardiovascular syndrome and risk of dementia in Japanese-American elderly men. The Honolulu-Asia aging study. *Arteriosclerosis Thrombosis, and Vascular Biology, 20*, 2255–2260.

Knecht, S., Breitenstein, C., Bushuven, S., Wailke, S., Kamping, S., Floel, A., et al. (2004). Levodopa: faster and better word learning in normal humans. *Annals of Neurology, 56*, 20–26.

Koch, G., Johansson, U., & Arvidsson, E. (1980). Radioenzymatic determination of epinephrine, norepinephrine and dopamine in 0.1 ml plasma samples: plasma catecholamine response to submaximal and near maximal exercise. *Journal of Clinical Chemistry and Clinical Biochemistry, 18*, 367–372.

Kraemer, W. J., Hakkinen, K., Newton, R. U., Nindl, B. C., Volek, J. S., McCormick, M., et al. (1999). Effects of heavy-resistance training on hormonal response patterns in younger vs. older men. *Journal of Applied Physiology, 87*, 982–992.

Krohne, H. W., Egloff, B., Kohlmann, C. W., & Tausch, A. (1996). Untersuchungen mit einer deutschen Version der "Positive and Negative Affect Schedule" (PANAS) [Investigations with a German version of the Positive and Negative Affect Schedule (PANAS)]. *Diagnostica, 42*, 139–156.

Larson, E. B., Wang, L., Bowen, J. D., McCormick, W. C., Teri, L., Crane, P., et al. (2006). Exercise is associated with reduced risk for incident dementia among persons 65 years of age and older. *Annals of Internal Medicine, 144*, 73–81.

Laurin, D., Verreault, R., Lindsay, J., MacPherson, K., & Rockwood, K. (2001). Physical activity and risk of cognitive impairment and dementia in elderly persons. *Archives of Neurology, 58*, 498–504.

Liang, K. C., Chen, L. L., & Huang, T. E. (1995). The role of amygdala norepinephrine in memory formation: involvement in the memory enhancing effect of peripheral epinephrine. *Chinease Journal of Physiology, 38*, 81–91.

Magnie, M. N., Bermon, S., Martin, F., Madany-Lounis, M., Suisse, G., Muhammad, W., et al. (2000). P300, N400, aerobic fitness, and maximal aerobic exercise. *Psychophysiology, 37*, 369–377.

Marie, R. M., & Defer, G. L. (2003). Working memory and dopamine: clinical and experimental clues. *Current Opinion in Neurology, 16*(Suppl 2), S29–S35.

McAllister, A. K., Katz, L. C., & Lo, D. C. (1999). Neurotrophins and synaptic plasticity. *Annual Review of Neuroscience, 22*, 295–318.

McMorris, T., & Graydon, J. (1997). The effect of exercise on cognitive performance in soccer-specific tests. *Journal of Sports Science, 15*, 459–468.

Meeusen, R., & De Meirleir, K. (1995). Exercise and brain neurotransmission. *Sports Medicine, 20*, 160–188.

Miyashita, T., & Williams, C. L. (2004). Peripheral arousal-related hormones modulate norepinephrine release in the hippocampus via influences on brainstem nuclei. *Behavioural Brain Research, 153*, 87–95.

Miyashita, T., & Williams, C. L. (2006). Epinephrine administration increases neural impulses propagated along the vagus nerve: role of peripheral beta-adrenergic receptors. *Neurobiology of Learning and Memory, 85*, 116–124.

Molteni, R., Ying, Z., & Gomez-Pinilla, F. (2002). Differential effects of acute and chronic exercise on plasticity-related genes in the rat hippocampus revealed by microarray. *European Journal of Neuroscience, 16*, 1107–1116.

Musso, N. R., Gianrossi, R., Pende, A., Vergassola, C., & Lotti, G. (1990). Plasma dopamine response to sympathetic activation in man: a biphasic pattern. *Life Science, 47*, 619–626.

Nakamura, Y., Nishimoto, K., Akamatu, M., Takahashi, M., & Maruyama, A. (1999). The effect of jogging on P300 event related potentials. *Electromyography and Clinical Neurophysiology, 39*, 71–74.

Neeper, S. A., Gomez-Pinilla, F., Choi, J., & Cotman, C. (1995). Exercise and brain neurotrophins. *Nature, 373*, 109.

Neeper, S. A., Gomez-Pinilla, F., Choi, J., & Cotman, C. W. (1996). Physical activity increases mRNA for brain-derived neurotrophic factor and nerve growth factor in rat brain. *Brain Research, 726*, 49–56.

Nutt, J. G., & Fellman, J. H. (1984). Pharmacokinetics of levodopa. *Clinical Neuropharmacology, 7*, 35–49.

Nybo, L., Nielsen, B., Blomstrand, E., Moller, K., & Secher, N. (2003). Neurohumoral responses during prolonged exercise in humans. *Journal of Applied Physiology, 95*, 1125–1131.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia, 9*, 97–113.

Pan, W., Banks, W. A., Fasold, M. B., Bluth, J., & Kastin, A. J. (1998). Transport of brain-derived neurotrophic factor across the blood–brain barrier. *Neuropharmacology, 37*, 1553–1561.

Pang, P. T., Teng, H. K., Zaitsev, E., Woo, N. T., Sakata, K., Zhen, S., et al. (2004). Cleavage of proBDNF by tPA/plasmin is essential for long-term hippocampal plasticity. *Science, 306*, 487–491.

Sakane, T., & Pardridge, W. M. (1997). Carboxyl-directed pegylation of brain-derived neurotrophic factor markedly reduces systemic clearance with minimal loss of biologic activity. *Pharmacal Research, 14*, 1085–1091.

Schinder, A. F., & Poo, M. (2000). The neurotrophin hypothesis for synaptic plasticity. *Trends in Neuroscience, 23*, 639–645.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron, 36*, 241–263.

Sharot, T., & Phelps, E. A. (2004). How arousal modulates memory: disentangling the effects of attention and retention. *Cognitive Affective & Behavioral Neuroscience*, 294–306.

Sibley, B. A., & Etnier, J. L. (2003). The relationship between physical activity and cognition in children: a meta-analysis. *Pediatric Exercise Science, 15*, 243–256.

Sjoberg, H. (1980). Physical fitness and mental performance during and after work. *Ergonomics, 23*, 977–985.

Sothmann, M. S., Gustafson, A. B., & Chandler, M. (1987). Plasma free and sulfoconjugated catecholamine responses to varying exercise intensity. *Journal of Applied Physiology, 63*, 654–658.

Spurway, N. C. (1992). Aerobic exercise, anaerobic exercise and the lactate threshold. *British Medical Bulletin, 48*, 569–591.

Sutoo, D., & Akiyama, K. (2003). Regulation of brain function by exercise. *Neurobiology of Disease, 13*, 1–14.

Tantillo, M., Kesick, C. M., Hynd, G. W., & Dishman, R. K. (2002). The effects of exercise on children with attention-deficit hyperactivity disorder. *Medicine and Science in Sports and Exercise, 34*, 203–212.

Tomporowski, P. D. (2003). Effects of acute bouts of exercise on cognition. *Acta Psychologica, 112*, 297–324.

Tomporowski, P. D., & Ellis, N. R. (1986). Effects of exercise on cognitive processes: a review. *Psychological Bulletin, 99*, 338–346.

Tomporowski, P. D., Ellis, N. R., & Stephens, R. (1987). The immediate effects of strenuous exercise on free-recall memory. *Ergonomics, 30*, 121–129.

van Gelder, B. M., Tijhuis, M. A., Kalmijn, S., Giampaoli, S., Nissinen, A., & Kromhout, D. (2004). Physical activity in relation to cognitive decline in elderly men: the FINE study. *Neurology, 63*, 2316–2321.

van Praag, H., Christie, B. R., Sejnowski, T. J., & Gage, F. H. (1999). Running enhances neurogenesis, learning, and long-term potentiation in mice. *Proceedings of the National Academy of Sciences of the United States of America, 96*, 13427–13431.

van Praag, H., Kempermann, G., & Gage, F. H. (1999). Running increases cell proliferation and neurogenesis in the adult mouse dentate gyrus. *Nature Neuroscience, 2*, 266–270.

Vaynman, S., Ying, Z., & Gomez-Pinilla, F. (2004). Hippocampal BDNF mediates the efficacy of exercise on synaptic plasticity and cognition. *European Journal of Neuroscience, 20*, 2580–2590.

Wang, G. J., Volkow, N. D., Fowler, J. S., Franceschi, D., Logan, J., Pappas, N. R., et al. (2000). PET studies of the effects of aerobic exercise on human striatal dopamine release. *Journal of Nuclear Medicine, 41*, 1352–1356.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology, 54*, 1063–1070.

Weuve, J., Kang, J. H., Manson, J. E., Breteler, M. M., Ware, J. H., & Grodstein, F. (2004). Physical activity, including walking, and cognitive function in older women. *JAMA: The Journal of the American Medical Association, 292*, 1454–1461.

Wise, R. A. (2004). Dopamine, learning and motivation. *Nature Reviews. Neuroscience, 5*, 483–494.

Wu, D. (2005). Neuroprotection in experimental stroke with targeted neurotrophins. *NeuroRx, 2*, 120–128.