Supplementary Materials for

Retrieval practice protects memory against acute stress
Amy M. Smith*, Victoria A. Floerke, Ayanna K. Thomas

*Corresponding author. Email: amy.smith@tufts.edu

Published 25 November 2016, Science 354, 1046 (2016)
DOI: 10.1126/science.aah5067

This PDF file includes:

Materials and Methods
Supplementary Text
Tables S1 to S3
References
Materials and Methods

Design
The experiment employed a 2 (learning strategy: retrieval practice vs. study practice) x 2 (TSST-G group: stressed vs. non-stressed) between-subjects factorial design.

Participants
One hundred twenty Tufts University students participated in the experiment (72 women, $M$ age = 20.08, $SD$ age = 3.95). Some participants were recruited through introductory psychology courses to fulfill a research participation requirement, and some were recruited from across the Tufts campus and received $20 for their participation. Thirty participants were randomly assigned to each of four groups: non-stressed study practice (SP), non-stressed retrieval practice (RP), stressed SP, and stressed RP. All participants provided informed consent.

Materials: Stimuli
Stimuli consisted of 30 nouns presented as words (15 neutral, 15 negative) and 30 nouns presented as images (15 neutral, 15 negative). All 60 stimuli were chosen from the Snodgrass and Vanderwart (28) pictures that have been normed according to emotional valence and arousal. The words presented were the nouns associated with each of 30 images. All images and words were semantically distinct.

Materials: State-Trait Inventory for Cognitive and Somatic Anxiety (STICSA)
We administered the STICSA (29) to assess participants’ self-reported levels of pre- and post-stress anxiety. Higher STICSA scores are indicative of higher self-reported anxiety.

Materials: The Trier Social Stress Test for Groups (TSST-G) Testing Room
To induce stress, we employed a modified version of the Trier Social Stress Test that accommodates group testing. The procedure for administering the TSST-G mimicked (18). In the testing room, participants were seated at four partitioned desks that were numbered 1-4 to facilitate participants being called on by the experimenter. When called on, participants stood and faced the front of the room where the two experimenters stood and took notes using clipboards. A camcorder was mounted on a tripod to the left of the experimenters and the camcorder appeared to be recording as they gave speeches and solved math problems.

Materials: Empatica E4 Wristbands
Because the TSST-G has been shown to reliably increase post-stress cortisol, we did not measure cortisol in the present study. However, we did measure heart rate during day 2 testing. Participants wore Empatica E4 wristbands (see www.empatica.com) for the duration of the day 2 experiment. The E4 is designed to measure blood volume pulse and interbeat intervals (IBI) in real-time. The E4 features a large button on the watch face, which participants were instructed to press at various points throughout the experiment to mark the onset and offset of different phases of the procedure.
Procedure: Stressed group.

Testing sessions occurred on two consecutive days between 3:30 p.m. and 5:30 p.m. to control for variability in diurnal cortisol secretion. We avoided morning testing because cortisol levels are naturally elevated in the morning (30), which could influence memory performance (e.g., 9).

Four participants partook in the experiment per session. On day 1, participants first began the encoding task, which was presented using E-Prime software (Version 2.1; 31). Participants were instructed that they would see several words and images, and that they would be given 10 seconds per item to type a sentence that included each given item. They were then presented with either the 30 words or 30 images at a rate of 2 seconds per item. Whether words or images were studied first was counterbalanced. After each item was presented, participants had 10 seconds to enter a sentence before the program advanced to the next item. Next, participants completed simple multiplication and division problems using pen and paper for 1 minute (e.g., 15 x 7). Those in the SP group then restudied the 30 items at a rate of 4 seconds per item, whereas those in the RP group were given 2 minutes to freely recall as many items as they could remember. On this test and all subsequent recall tests in this experiment, participants were not given feedback as to the correctness of their answers. This procedure (item presentation and sentence generation, math, and restudy or free recall) was then repeated for the 30 items of the other item-type. During a subsequent 5-minute retention interval, participants worked on a SUDOKU puzzle. SP participants then engaged in re-study of all 30 words followed by all 30 images at a rate of 3 seconds per item, whereas RP participants were given 3 minutes for free recall. During free recall, RP participants were instructed to type as many words and images as they could remember and to record a “W” next to items that had been presented as words and an “I” next to items that had been presented as images. Another 5-minute retention interval then followed, in which participants worked on a new SUDOKU puzzle. SP participants then engaged in a final round of re-study (60 items, 3 seconds per item) and RP participants were given 3 minutes for free recall.

After encoding, participants filled out the first iteration of the STICSA. Participants were then excused and reminded to return the next day for the second part of the experiment. Those who received payment for their participation were given $10 at the end of day 1. The day 1 experimental procedure lasted approximately 45 minutes.

On day 2, participants returned to the original testing room 24 hours after the first session. The undergraduate experimenter first asked them to complete the STICSA for a second time. She then fastened an Empatica E4 wristband around each participant’s non-dominant wrist and instructed participants to sit still for one minute while the devices collected a baseline measure of heart rate. A graduate student experimenter then entered the room, dressed in business attire. She gave each participant a blank sheet of paper, and instructed them to prepare a speech in which they would be applying for the position of a teaching assistant in a class of their choice. After 2 minutes, an experimenter abruptly took participants’ notes away and instructed them that they would give their speeches extemporaneously. The graduate experimenter turned on the video camera and told
participants that they would be video recorded during their speeches for the purpose of coding their non-verbal behavior at a later time. The graduate experimenter then called on participants in random order to deliver 1-minute speeches. During each speech, the experimenters took notes on a clipboard and withheld verbal and non-verbal feedback.

After giving their speeches, participants were given test 1, a pen-and-paper free recall test in which they were asked to recall either the words or the images that they had learned the previous day. Initial recall of words or images was counterbalanced. Participants were given 2.5 minutes for test 1.

Participants were next called on randomly to orally subtract numbers in the teens from four-digit numbers (e.g., 4,866 - 19). Each participant was called on multiple times during the 6-minute subtraction phase. The TSST-G stress induction took approximately 15 minutes, including the speech preparation, individual speeches, recall test 1, and subtraction task.

After the math portion of the TSST-G, participants completed the STICSA for the third and final time. During a subsequent 10-minute resting period leading up to the final memory test, participants viewed part of an episode of the NBC television series The Office. Afterward, an experimenter gave participants recall test 2, prompting them to remember items of the item-type (words or images) that had not been assessed on test 1. After 2.5 minutes had passed, the experimenter collected the tests.

Following the second memory test, participants were paid (when applicable) and debriefed. The experimenter explained to participants that they had not actually been videotaped and that the experimenters were not judging their non-verbal behavior. The day 2 experimental procedure lasted approximately 45 minutes.

Procedure: Non-stressed group

The procedure for the non-stressed group followed the same protocol as discussed for the stress group, with the exception that the non-stressed group did not receive the TSST-G stress manipulation on day 2. The TSST-G protocol for non-stressed participants followed (18) and was designed to mimic the procedure for stressed participants without the components of socio-evaluative threat and unpredictability. In place of the 2-minute speech preparation phase and the 4-minute speech phase, participants in the non-stressed group sat and silently read a chapter from a biology textbook for 6 minutes. In place of the 6-minute oral math subtraction task, participants in the non-stressed group solved math subtraction problems with pen-and-paper for 6 minutes. They were given as much time as they needed to complete each problem and were told that their answers would not be graded. The two experimenters were present in the room during these tasks but wore casual clothing and did not observe or question the participants. All other experimental procedures were identical to those of the stress group.
Statistical Analysis

Physiological Arousal

To examine whether the TSST-G induced a physiological stress response, we examined heart rate variability as measured by blood volume pulse and IBI. Blood volume pulse was measured in number of beats per minute, and IBI was measured in number of milliseconds between heart beats. We used the MATLAB Kubios software package (see http://kubios.uef.fi/) to calculate each participant’s average pulse and IBI over the span of the 1-minute baseline measurement and over the span of the 12-minute TSST-G task. The 12-minute TSST-G measurement did not include the 2.5-minute memory test (test 1) that occurred in the middle of the TSST-G on day 2, since the test was not part of the stress manipulation.

One Empatica E4 wristband was unpredictably faulty, resulting in physiological data for only 104 of 120 participants. Thus, the following analyses were conducted on 58 participants who completed the non-stressed TSST-G tasks, and 46 participants who completed the stress-induction tasks.

We ran paired-samples t tests comparing mean pulse and IBI during the stressed and non-stressed TSST-G tasks to mean activity during the baseline measurement on day 2. As expected, stressed participants demonstrated post-stress increases in pulse (t(45) = 7.53, p < .001, d = 0.82) and decreases in IBI (t(45) = 5.05, p < .001, d = 0.38) relative to baseline, whereas non-stressed participants did not show changes in pulse (t(57) = 1.77, p = .08) or IBI (t(57) = 0.21, p = .84). Pulse and IBI averages are reported in Table 1.

Self-reported Stress

Because the act of taking tests may be stressful for some participants, we first examined whether our day 1 manipulation (retrieval practice vs. study practice) affected participants’ subsequent self-reported levels of stress. An independent samples t test revealed no difference in STICSA scores for participants who had engaged in retrieval practice versus study practice (t(118) = 0.56, p = .58).

To test whether the TSST-G tasks increased subjective anxiety on day 2, we ran paired-samples t tests comparing pre- and post-TSST-G STICSA scores. As expected, stressed participants demonstrated heightened post-stress STICSA scores relative to baseline (t(59) = 3.30, p < .001, d = 0.27), whereas non-stressed participants did not (t(59) = 0.54, p = .60). STICSA averages are reported in Table 1.

Day 1 Memory Performance

Table 2 displays correct recall averages for the participants who were given the RP manipulation on day 1. On the final two recall tests (T2 and T3) on day 1, participants were instructed to recall both words and images and indicate the source from which each item came (i.e., the word list or the image list). On average, participants made 0.4 (SEM = 0.09) source misattributions on T2 and 0.4 (SEM = 0.10) source misattributions on T3. Because day 1 memory performance was not relevant to the questions posed by the present study, we did not examine it any further.
Day 2 Memory Performance: Approach to Data Analysis

In our analyses on day 2 memory performance, we did not examine memory for words and images separately. To control for picture superiority effects, we counterbalanced whether words or images were recalled during the first memory test, with the other item-type being recalled during the second test. Because words and images were recalled by an equal number of participants on each memory test, we collapsed across item type when examining means.

Furthermore, gender differences have been reported in both the stress and memory literature (e.g., 2, 6), as well as the broader memory literature (e.g., 32, 33), with women outperforming men on tests of verbal memory. To control for these gender differences, we included gender as a covariate in all of the following analyses.

Day 2: Recall During the Immediate Stress Response

We conducted a two-way analysis of covariance (ANCOVA) to determine the effects of learning strategy (RP vs. SP) and TSST-G group (stressed vs. non-stressed) on test 1 recall, controlling for gender. We found a main effect of learning strategy, as those who learned via RP recalled more items than those who learned via SP (9.88 vs. 8.15 items), $F(1, 115) = 7.60, p < .01, \eta^2_p = .06$. Confirming our prediction that gender would act as a covariate, we also found a significant main effect of gender, with females recalling more items on average than males (9.58 vs. 8.17 items), $F(1, 115) = 4.12, p < .05, \eta^2_p = .04$. All other effects, including the main effect of the TSST-G manipulation, were non-significant (all $p$’s > .50). Average test 1 recall is displayed in Fig. 1 and Table 1.

Day 2: Recall During the Delayed Stress Response

We first conducted a planned independent samples $t$ test comparing test 2 recall performance for stressed versus non-stressed participants, all of whom learned via SP. This analysis served to replicate the several previous studies that found impaired memory performance for stressed participants when retrieval was assessed after a post-stress delay. Indeed, stressed SP participants recalled significantly fewer items than non-stressed SP participants on test 2, $t(58) = 1.71, p < .05, d = 0.44$.

We next conducted a two-way ANCOVA to determine the effects of learning strategy (RP vs. SP) and TSST-G group (stressed vs. non-stressed) on test 2 recall, again controlling for gender. Most notably, we found a significant learning strategy by TSST-G group interaction, $F(1, 115) = 4.24, p < .05, \eta^2_p = .04$. We investigated the nature of this interaction using independent samples $t$ tests, comparing $p$ to an alpha value of .008 (.05/6). As Fig. 1 shows, both stressed and non-stressed RP participants recalled significantly more items than stressed SP participants (respectively, $t(58) = 4.50, p < .001, d = 1.16$; $t(58) = 3.77, p < .001, d = 0.97$). This analysis also found a main effect of learning strategy, as those who learned via RP recalled more items than those who learned via SP (10.68 vs. 7.88 items), $F(1, 115) = 16.65, p < .001, \eta^2_p = .13$. Lastly, once again confirming the influence of gender as a covariate, we found a marginally significant main effect of gender as females recalled more items than males (9.86 vs. 8.42...
items), $F (1, 115) = 3.20, p = .08, \eta_p^2 = .03$. Average test 2 recall is displayed in Fig. 1 and Table 1.

**Day 2: Source Misattributions**

Table 5 displays average source misattributions on test 1 and test 2. We ran two exploratory two-way ANCOVAs (controlling for gender) to determine whether learning strategy (RP vs. SP) and TSST-G group (stressed vs. non-stressed) affected participants’ source misattributions on test 1 and test 2. All main effects and interactions were non-significant (all $p$’s > .20).
Table S1.
Average STICSA scores, pulse, and IBI as a function of TSST-G group (stressed vs. non-stressed). Standard errors of the mean are given in parentheses.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Non-stressed Baseline</th>
<th>Non-stressed During TSST-G Task</th>
<th>Stressed Baseline</th>
<th>Stressed During TSST-G Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>STICSA Score</td>
<td>28.6 (0.83)</td>
<td>29.0 (0.95)</td>
<td>31.6 (1.13)</td>
<td>33.9 (1.08)</td>
</tr>
<tr>
<td>Pulse (bpm)</td>
<td>75.0 (1.55)</td>
<td>75.9 (1.69)</td>
<td>73.1 (1.93)</td>
<td>84.7 (2.11)</td>
</tr>
<tr>
<td>IBI (ms)</td>
<td>883.7 (21.05)</td>
<td>889.7 (22.13)</td>
<td>929.7 (25.67)</td>
<td>866.4 (22.18)</td>
</tr>
</tbody>
</table>
Table S2.
Average number of items correctly recalled on day 1 for participants given the RP manipulation (participants given the SP manipulation did not engage in recall on day 1). Standard errors of the mean are given in parentheses. Note: Words and images were tested in separate blocks on the first test (T1). On the last two tests (T2 and T3), participants recalled words and images together (see Materials and Methods for more detail).

<table>
<thead>
<tr>
<th></th>
<th>T1</th>
<th></th>
<th>T2</th>
<th></th>
<th>T3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Words</td>
<td>Images</td>
<td>Words</td>
<td>Images</td>
<td>Words</td>
<td>Images</td>
</tr>
<tr>
<td>Negative</td>
<td>6.6 (0.26)</td>
<td>6.9 (0.32)</td>
<td>4.8 (0.25)</td>
<td>6.1 (0.28)</td>
<td>5.3 (0.27)</td>
<td>6.4 (0.29)</td>
</tr>
<tr>
<td>Neutral</td>
<td>5.9 (0.30)</td>
<td>5.7 (0.24)</td>
<td>4.3 (0.29)</td>
<td>5.0 (0.29)</td>
<td>4.5 (0.28)</td>
<td>5.7 (0.30)</td>
</tr>
<tr>
<td>Total</td>
<td>12.5 (0.43)</td>
<td>12.4 (0.47)</td>
<td>9.0 (0.44)</td>
<td>11.0 (0.47)</td>
<td>9.8 (0.46)</td>
<td>12.1 (0.49)</td>
</tr>
</tbody>
</table>
Table S3.
Average number of source misattributions on test 1 and test 2 on day 2. Standard errors of the mean are given in parentheses.

<table>
<thead>
<tr>
<th>Group</th>
<th>Test 1</th>
<th>Test 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-stressed</td>
<td>1.3 (0.49)</td>
<td>0.8 (0.21)</td>
</tr>
<tr>
<td>Stressed</td>
<td>1.4 (0.39)</td>
<td>0.9 (0.29)</td>
</tr>
<tr>
<td>RP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-stressed</td>
<td>1.1 (0.48)</td>
<td>0.7 (0.32)</td>
</tr>
<tr>
<td>Stressed</td>
<td>0.7 (0.25)</td>
<td>0.3 (0.12)</td>
</tr>
</tbody>
</table>
References and Notes


19. Materials and methods are available as supplementary materials on *Science* Online.


