American Psychological Association
Behavioral Neuroscience

0735-7044/18/$12.00
2018, Vol. 132, No. 3, 161–170
http://dx.doi.org/10.1037/bne0000240

Criterial Learning is Not Enough: Retrieval Practice is Necessary for Improving Post-Stress Memory Accessibility

Amy M. Smith
Tufts University

F. Caroline Davis
U.S. Army Natick Soldier Research, Development, and Engineering Center (NSRDEC), Natick, Massachusetts and Tufts University

Ayanna K. Thomas
Tufts University

In a recent study, having participants make three retrieval attempts (i.e., retrieval practice) when learning information strengthened memory against the detrimental effects of psychological stress. We aimed to determine whether learning to criterion, in which only one successful retrieval attempt is made, would similarly buffer memory against stress, or whether multiple retrieval attempts are necessary to achieve that effect. In Experiment 1, participants learned to criterion and then engaged in additional restudying (CL_S) or retrieval practice (CL_R). Twenty-four hours later, stress was induced and stress-related increases in cortisol were observed. However, no differences in recall performance were observed between any of the groups. Experiment 2 was similar but introduced a 1-week delay between encoding and retrieval. Recall performance was impaired for both groups under stress, but recall for those in the CL_R group was still better than either pre- or post-stress performance for those in the CL_S group. Thus, criterial learning may protect memory against stress in the short-term, but additional retrieval practice is more beneficial for achieving this effect in the long-term.

Keywords: stress, retrieval practice, testing effect, memory, cortisol

Instances of acute psychological stress, such as having to give a speech or take an exam, have been shown to impair memory retrieval (Shields, Sazma, McCullough, & Yonelinas, 2017). This robust finding has far-reaching implications; in high-pressure, high-stakes scenarios, access to crucial memories may be blocked. However, recent findings suggest that taking multiple practice memory tests while studying, a technique referred to as *retrieval practice*, can improve post-stress memory accessibility (Smith, Floerke, & Thomas, 2016). The present research examined whether successfully retrieving information once can create stress-resistant memories, or whether additional retrieval practice is necessary. In two experiments, we examined this relationship across 24-hr (Experiment 1) and 1-week (Experiment 2) retention intervals.

Implicated in stress-related memory impairment is the human stress response, during which catecholamines and cortisol are released to prepare the body to take defensive action. While memory retrieval may be enhanced (Hupbach & Fieman, 2012; Schönfeld, Ackermann, & Schwabe, 2014) or unaffected (Schwabe & Wolf, 2014) by the immediate surge of catecholamines, retrieval is generally impaired once cortisol reaches peak post-stress levels in the blood 20–30 min later (see Gagnon & Wagner, 2016; Schwabe, Joels, Roozendaal, Wolf, & Oitzl, 2012; Shields et al., 2017). Via its occupation of glucocorticoid receptors in the amygdala and hippocampus (Lovallo, Robinson, Glahn, & Fox, 2010; Reul & de Kloet, 1985) and the resulting suppression of activity in prefrontal regions (Gärtner, Rohde-Liebenau, Grimm, & Bajbouj, 2014; Qin, Hermans, van Marle, Luo, & Fernández, 2009), cortisol effectively prepares the brain for the learning and consolidation of information at the cost of suppressing processing in these retrieval-related areas (Schwabe et al., 2012). Further, post-stress memory impairment has been shown to be more pronounced for stimuli of negative or positive valence than for neutral stimuli (see Shields et al., 2017), likely because of the influence of cortisol on the amygdala.

Though most studies examining the effects of stress on retrieval have found detrimental effects, a few studies lend insight into the conditions under which stress may not impair memory. In a standard stress-and-memory paradigm, participants study stimuli and return 24 hr later for stress induction and subsequent memory testing. Deviating from this norm, some researchers had participants take one or more memory tests immediately after studying

---

Amy M. Smith, Department of Psychology, Tufts University; F. Caroline Davis, U.S. Army Natick Soldier Research, Development, and Engineering Center (NSRDEC), Natick, Massachusetts, and Center for Applied Brain and Cognitive Sciences, Tufts University; Ayanna K. Thomas, Department of Psychology, Tufts University.

This work was supported by the U.S. Army Natick Soldier Research, Development, and Engineering Center (NSRDEC). Aspects of this research were presented at the annual meetings of the Psychonomic Society and the Association for Psychological Science in 2017.

Correspondence concerning this article should be addressed to Amy M. Smith, Department of Psychology, Tufts University, 490 Boston Avenue, Medford, MA 02155. E-mail: amy.smith@tufts.edu

stimuli on the first day of the experiment (e.g., Beckner, Tucker, Delville, & Mohr, 2006; Buchanan & Tranel, 2008; Oei, Everaerd, Elzinga, van Well, & Bermond, 2006; Schoofs & Wolf, 2009; Wolf, Schommer, Hellhammer, Reischies, & Kirschbaum, 2002). In general, retrieval practice of this nature yields better long-term memory accessibility than conventional forms of studying such as rote rehearsal (see Roediger & Butler, 2011; Roediger & Karpicke, 2006). Thus, the purpose of using retrieval practice in previous stress studies was to ensure memory retrievability after a 24-hr delay. Considering the efficacy of retrieval practice, it is unsurprising that two of these studies did not find stress-related retrieval impairment (Schoofs & Wolf, 2009; Wolf et al., 2002). The use of retrieval practice in those experiments may have bolstered memory against stress, resulting in the observed null effects.

In a recent study, we directly tested this hypothesis (Smith et al., 2016). After all participants initially studied stimuli, half of them restudied the stimuli three times and half completed three recall tests. Twenty-four hours later, participants underwent stress induction or a time-matched control task, and then completed a recall test. Replicating the many previous studies, among individuals who engaged in restudying, those who were stressed recalled fewer items than those who were not stressed. However, for those who engaged in retrieval practice, stressed and nonstressed participants demonstrated similar recall and outperformed both restudy groups. Thus, when retrieval practice was directly manipulated to promote memory accessibility, the typical impairing effects of stress were no longer observed.

This study suggests that making multiple retrieval attempts during learning may improve post-stress memory accessibility. However, it remains to be determined whether multiple attempts are necessary to achieve this or whether one retrieval attempt is sufficient. In general, there is a curvilinear relationship between the number of retrieval practice attempts that are made during encoding and the number of items remembered on a later test (Pyc & Rawson, 2009). That is, the first few retrieval attempts yield large memory improvements, with diminishing returns thereafter. An interesting find was that both self-report (Kornell & Bjork, 2007; Wissman, Rawson, & Pyc, 2012) and experimental (Kornell & Bjork, 2008) evidence suggests that most students who use self-testing (e.g., flashcards) will only test themselves until they can accurately recall each desired item once. Students assume that learning to a criterion of one (i.e., retrieving an item once) is sufficient for remembering information on a future test. Thus, determining whether one successful retrieval attempt while studying can buffer memory against stress to the same degree as three retrieval attempts would provide useful information for learners who already default to a criterial learning strategy. Further, knowing whether criterial learning effectively creates stress-resistant memories would help optimize the time that individuals spend studying in preparation for a stressful event. Finally, by characterizing the boundary between initial learning and subsequent strengthening through practice, we can better understand the relationship between the underlying memory processes and the physiological processes associated with an acute stress response.

The present research examined whether learning to a criterion of one could bolster memory against stress, or whether additional retrieval practice may be necessary to achieve that goal. In two experiments, we had participants engage in a criterial learning task followed by either additional restudying or retrieval practice. Thus, participants successfully recalled information one time and subsequently engaged in additional studying or retrieval practice. After a 24-hr (Experiment 1) or 1-week (Experiment 2) delay, participants returned to the lab for stress induction followed by cued recall testing. The key difference between the two experiments was the retention interval between encoding and post-stress retrieval, which was manipulated because the advantages of retrieval practice over restudying have been shown to become increasingly pronounced as a function of time (e.g., Thompson, Wenger, & Bartling, 1978; Wheeler, Ewers, & Buonanno, 2003). In Experiment 1, we tested whether the benefits of additional, postcriterial-learning retrieval practice would emerge after the typical 24-hr delay. In Experiment 2, we tested whether these benefits would be further exaggerated after a longer delay of 1 week.

Experiment 1

In Experiment 1, participants engaged in a criterial learning task followed by either additional restudying or retrieval practice. After a 24-hr delay, they returned to the lab for stress induction or a control task followed by cued recall testing. Recently, researchers have opted to test memory during both the first phase of the physiological stress response when catecholamines are released and immediately after the onset of stress during the cortisol peak (Schönfeld et al., 2014; Schwabe & Wolf, 2014; Smith et al., 2016). As discussed earlier in the article, these studies have begun to provide evidence for a positive or neutral influence of catecholamines on retrieval and a negative influence of cortisol on retrieval. In Experiment 1, we adopted this two-test paradigm to provide further support for this growing body of literature. As has been done in the previous two-test experiments, there was no overlap in the content tested on each of the two tests.

We modeled our criterial learning task after Karpicke and Roediger (2008), who found that participants who engaged in criterial learning followed by additional studying (CL_S) recalled 36% of stimuli 1 week later, whereas those for whom criterial learning was followed by additional retrieval practice (CL_R) recalled 80% of stimuli. Based on their findings we hypothesized that CL_R participants would outperform CL_S participants. Based on the findings of Smith et al. (2016), we further hypothesized that stressed and nonstressed CL_R participants would demonstrate similar memory performance, but that stressed CL_S participants would demonstrate lower cued recall than those in the nonstressed CL_S group.

Method

Design. The experiment used a 2 (learning strategy: CL_S or CL_R) × 2 (TSST-G group: nonstressed or stressed) between-subjects factorial design.

Participants. Assuming an effect size of $\eta_p^2 = .04$ derived from Smith et al. (2016), a significance level of $\alpha = .05$, and four between-subjects groups, an a priori power analysis suggested a sample size of 191 participants to achieve 80% power to detect effects (G*Power 3.0; Faul, Erdfelder, Lang, & Buchner, 2007). However, the experiment was terminated part-way through data collection when preliminary analyses revealed null effects backed by strong statistical support (see the Results section for more detail). Thus, 133 Tufts University students participated in the experiment. Twenty participants were excluded from data analysis for the following reasons: experimenter error ($n = 4$), prior knowledge of Swahili ($n = 1$), failure to return for the second experimental session ($n = 3$), or perfect memory performance ($n = 12$). As such, all final analyses were conducted on 113 participants (75 women, $M_{\text{age}} = 18.96$, $SD_{\text{age}} = 1.41$), all of whom reported that they had not consumed caffeine or nicotine in the 6 hr before the experiment. 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. Participants were excluded if they had previously participated in an experiment involving psychological stress induction. Twenty-eight participants were randomly assigned to the nonstressed CL_S group, 27 to the nonstressed CL_R group, 28 to the stressed CL_S group, and 30 to the stressed CL_R group. All participants provided informed consent.

Materials.

Stimuli. The learning materials consisted of 40 Swahili nouns of mixed valence/arousal, each paired with their English translation (e.g., *bustani—garden*). The stimuli were chosen because they had been used in a prior study on criterial learning and did not result in ceiling memory performance (Karpicke & Roediger, 2008).

Anxiety questionnaire. We administered the State-Trait Inventory for Cognitive and Somatic Anxiety (STICSA) to assess participants' self-reported levels of pre- and post-stress anxiety (Grös, Antony, Simms, & McCabe, 2007). STICSA scores range from 0–80 and higher scores are indicative of higher self-reported anxiety.

Procedure. 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 (Weitzman et al., 1971). Stressed participants were tested four at a time according to stress induction protocol, and nonstressed participants were tested in groups of 2–4.

On Day 1, participants first began the criterial learning task, which was presented using E-Prime software (Version 2.1; Schneider, Eschman, & Zuccolotto, 2001). Participants were instructed that they would learn a series of Swahili words by studying the words multiple times and taking multiple memory tests for the words. They then began the first of four study/test cycles, which was the same for all participants. Participants viewed 40 Swahili-English word pairs, presented individually and in random order for 5 s each. A 30 s retention interval followed, in which participants completed simple math problems (e.g., 12 × 9) that they were told would not be graded. Afterward, participants were presented with all 40 Swahili words and were asked to type each English translation. The Swahili words were presented individually and in alphabetical order. Participants were given 8 s to enter a response before the program advanced to the next Swahili cue word. No feedback regarding correctness was given.

Participants next completed three similar study/test cycles, which differed according to group. In the CL_S group, each of the three study phases consisted of restudying all 40 word pairs in random order. During each of the three test phases, those in the CL_S group were prompted to recall the English translations for the subset of Swahili words that they had not correctly recalled during any previous retrieval attempt. In cases in which these participants correctly recalled all 40 items before the fourth study/test cycle, no subsequent retrieval attempts were made and thus they engaged in back-to-back study periods. In the CL_R group, each of the three study phases consisted of restudying the subset of words that they had not recalled on any previous cued-recall test. During the three test phases, participants in the CL_R group were prompted to recall the English translation for each of the 40 items. When these participants correctly recalled all 40 items before the fourth study/test cycle, no subsequent restudying was required and, thus, they engaged in back-to-back retrieval attempts. In both groups, the program ended once all four study/test cycles had been completed. Participants then filled out the first STICSA, were paid (when applicable), and excused.

Twenty-four hours after learning, participants returned to the original testing room where they completed the second STICSA and provided a baseline saliva sample for cortisol analysis. Participants then began the procedure associated with either the control or stress version of the Trier Social Stress Test for Groups (TSST-G; von Dawans, Kirschbaum, & Heinrichs, 2011). Briefly, those in the stress groups were given 2 min to prepare a speech and then took turns giving their 2 min speeches extemporaneously while being videotaped and judged. Participants in the time-matched control groups read silently from a biology textbook. After the speech or reading task, participants provided the second saliva sample.

All participants were next given Test 1 to assess memory performance during the immediate stress response. They were presented with a list of 20 of the 40 Swahili cue words on a single sheet of paper and were given 2.5 min to write the English translations. Participants were not forced to answer each question, and were permitted to change their responses and answer items in any order. All participants completed the test within the allotted time.

The TSST-G procedure then continued, as participants in the stress groups were called on at random 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-min subtraction phase. Participants in the control groups were given as much time as needed to solve the same math problems using pen and paper. After the TSST-G, participants completed the third STICSA.

Next, during a 10-min resting period leading up to the final memory test, participants viewed part of an episode of the NBC TV series *The Office*. They then provided the third saliva sample to assess peak post-stress levels of cortisol. Participants were then given 2.5 min to complete Test 2, which was identical to Test 1 but assessed participants' memory for the 20 Swahili words that were not previously tested. Finally, participants were paid (when applicable), debriefed, and excused. We refer readers to Figure 1 for a graphic depiction of the Experiment 1 procedure.

Cortisol measurement and data management. Saliva samples were stored at −20 °C until the completion of data collection, after which they were shipped to Salimetrics, LLC (Salimetrics, LLC, State College, PA; Salimetrics, LLC, 2017) for analysis. Samples were assayed in duplicate, and the mean cortisol concentration served as the dependent measure. We converted cortisol concentrations from micrograms per deciliter to nanomoles per liter for consistency with the majority of human stress literature.

Results

Self-reported stress. Because the act of taking tests may be stressful for some participants, we first examined whether our Day 1 manipulation (CL_S vs. CL_R) affected participants' subsequent

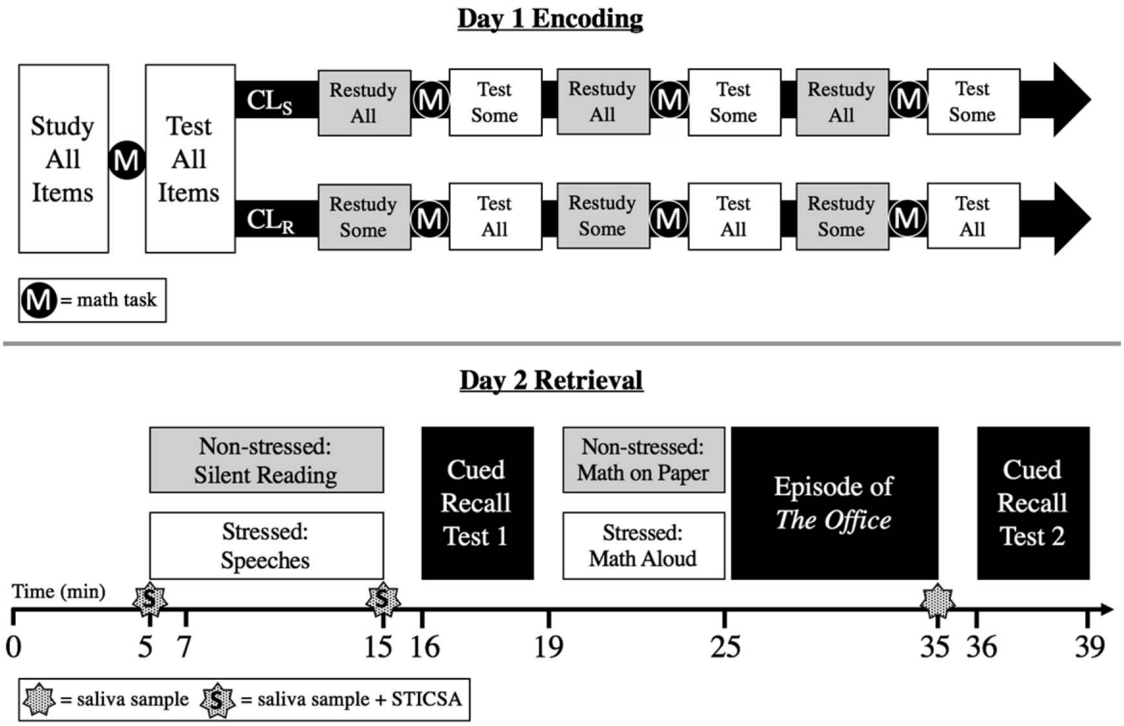

Figure 1. Graphic depiction of the Experiment 1 procedure. In Experiment 2, the Day 1 procedure is identical to that used in Experiment 1, whereas the Day 2 tasks were slightly modified in Experiment 2 (see the Experiment 2 Method section). CL$_S$ denotes participants who engaged in study practice after criterial learning, and CL$_R$ denotes those who engaged in retrieval practice after criterial learning.

self-reported levels of stress. An independent samples *t* test on average Day 1 STICSA scores revealed no difference for participants who had engaged in additional retrieval practice versus additional study practice, $t(111) = 1.071$, $p = .286$.

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, $t(54) = 5.670$, $p < .001$, $d = 0.83$, whereas nonstressed participants did not, $t(54) = 0.841$, $p = .404$. Average Day 2 STICSA scores are shown in Figure 2.

Cortisol. We next examined cortisol reactivity to the TSST-G tasks. A 3 (time: baseline, 10 min post-stress, 30 min post-stress) × 2 (TSST-G group: stressed or nonstressed) mixed model analysis of variance (ANOVA) revealed a significant interaction, $F(2, 214) = 4.219$, $p = .016$, $\eta_p^2 = .038$. Pairwise comparisons using a Bonferroni correction showed that, as expected, stressed participants demonstrated increased cortisol from baseline to 30 min post-stress (mean difference = 2.852, *SEM* = .542, $p < .001$) and from 10 min post-stress to 30 min post-stress (mean difference = 2.367, *SEM* = .800, $p = .011$), whereas nonstressed participants demonstrated no differences in cortisol across the three measurements (*p*s > .10). Thus, the TSST-G manipulation successfully induced a physiological stress response for individuals in the stress groups. Average cortisol concentrations are shown in Figure 3.

Day 1 criterial learning. As shown in Figure 4, there were no differences in the learning curves for participants who engaged in additional study practice or additional retrieval practice during the criterial learning task on Day 1. This was confirmed by a 2 (learning strategy: CL$_S$, CL$_R$) × 4 (study/Test Trial: 1, 2, 3, 4) mixed ANOVA with study/test trial as a within-subjects variable, in which there was no main effect of learning strategy on how many items participants correctly remembered on each cued-recall test, $F(1, 111) = 0.105$, $p = .746$. By the end of the task, participants had accurately recalled an average of 35 out of 40 items at least once, or approximately 88%.

Day 2 memory performance. Because gender differences have been reported in the stress and memory literature (e.g., Buchanan & Tranel, 2008; Schwabe & Wolf, 2014; Smith et al., 2016) and the broader memory literature more generally (e.g., Herlitz, Nilsson, & Bäckman, 1997; Lewin, Wolgers, & Herlitz, 2001), we included gender as a covariate in all of the following omnibus analyses.

Test 1. We conducted a two-way analysis of covariance (ANCOVA) to determine the effects of learning strategy (CL$_S$ vs. CL$_R$) and TSST-G group (stressed vs. nonstressed) on Test 1 cued recall, controlling for gender. All main effects and the interaction were nonsignificant (all *F*s < 1.4). We then repeated this analysis after removing individuals who did not demonstrate a cortisol response to stress (i.e., 11 nonresponders). Thus, we eliminated eight nonresponders (seven women) in the stressed CL$_S$ group and three nonresponders (all women) in the stressed CL$_R$ group. All effects were again nonsignificant (all *F*s < 1.2). Average Test 1 cued recall performance is displayed in Table 1.

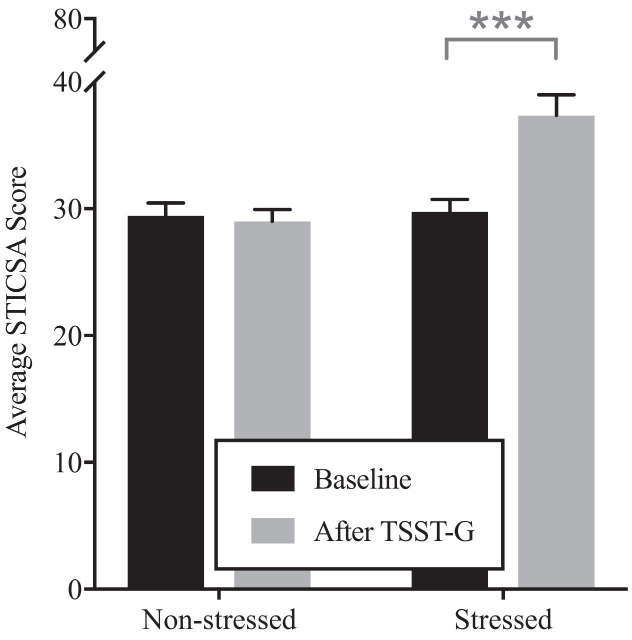

Figure 2. Average pre- and post-TSST-G (Trier Social Stress Test for Groups) STICSA (State-Trait Inventory for Cognitive and Somatic Anxiety) scores for stressed and nonstressed participants on Day 2 in Experiment 1. Error bars represent *SEM*. *** $p < .001$.

Test 2. We next conducted identical analyses on Test 2 cued recall. Again, all main effects and interactions were nonsignificant for the analysis on all participants and the analysis excluding nonresponders (all Fs < 1.2). Test 2 recall performance is also displayed in Table 1.

The null effect found for Test 2 performance stands in contrast to the Experiment 1 results, which featured a significant interaction between learning strategy and TSST-G group on the second memory test. To further support the null finding reported here, the Test 2 data were examined by estimating Bayes factors for both main effects and the interaction. Bayes factors have emerged as a way to support the conclusion that null results are sound evidence in favor of the null hypothesis and are not the result of data insensitivity (Dienes, 2014). Unlike frequentist approaches, such as power estimates that are based on preexisting values and assumptions (e.g., α values and estimated effect sizes), Bayes factors rely only on the observed data to determine whether evidence favors the null or alternative hypothesis. Bayes factors range from 0 to ∞, with values close to 0 being strong evidence in favor of the null hypothesis, values much greater than 1 being strong evidence in favor of the alternative hypothesis, and values around 1 suggesting that the data are insensitive (Dienes, 2014). Bayes factors were analyzed using the statistical software program JASP (2017), assuming the default recommended fixed effect and random effect prior values. The estimated Bayes factors (BF_{10}) for the effects of learning strategy, TSST-G group, and the interaction were 0.20, 0.20, and 0.31, respectively. All three values are considered strong evidence that the data are more likely to occur under the null than the alternative hypothesis (Jeffreys, 1939/1961), further supporting the null findings of the previous ANCOVA.

To summarize, in Experiment 1, our stress manipulation successfully induced subjective and physiological stress. Despite this, we found no differences in performance on either an immediate or delayed test of memory when comparing our four experimental groups: nonstressed CL_S, nonstressed CL_R, stressed CL_S, and stressed CL_R. These findings occurred in the context of a 24-hr delay that separated encoding and retrieval.

Experiment 2

In Experiment 1, all individuals performed similarly on the final cued-recall test, regardless of the learning strategy they used on Day 1 and whether stress was induced before retrieval on Day 2. Cued recall performance was generally high in Experiment 1, as evidenced by the number of participants whose data were excluded because of ceiling cued recall performance. As such, the question remains whether the benefits of additional retrieval practice after learning to criterion would emerge after a longer delay when more forgetting has occurred. In fact, in several studies, the benefits of retrieval practice over restudying did not emerge until 48 hr or even 1 week after initial learning (for a review see Roediger & Karpicke, 2006). In the context of criterial learning, in which all participants engage in retrieval practice at least once, the benefits of additional retrieval attempts may be particularly difficult to detect in the few days after initial learning. Thus, in Experiment 2, we examined whether learning to criterion would continue to protect memory against stress after a 1-week delay, or whether additional postcriterial-learning retrieval practice would provide a benefit over additional restudying.

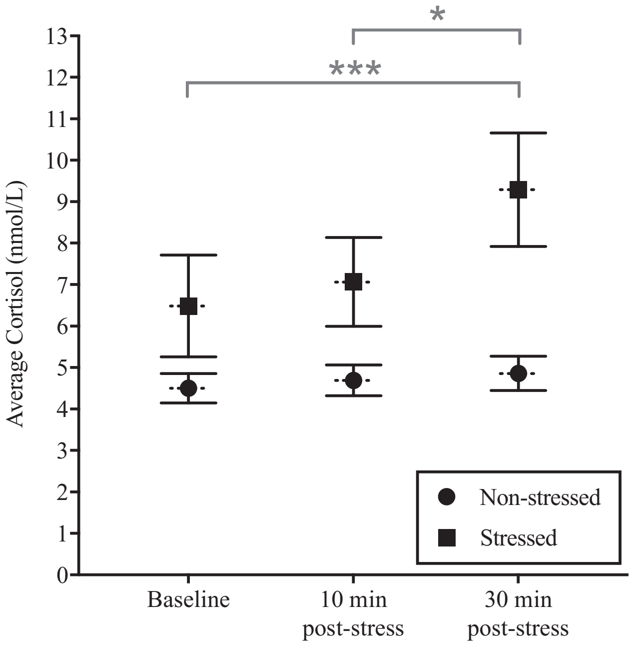

Figure 3. Average cortisol concentrations for stressed and nonstressed participants on Day 2 in Experiment 1. Cortisol samples were taken immediately before the TSST-G (baseline), 10 min after the onset of the TSST-G (Trier Social Stress Test for Groups), and 30 min after the onset of the TSST-G. Error bars represent *SEM*. * $p < .05$. *** $p < .001$.

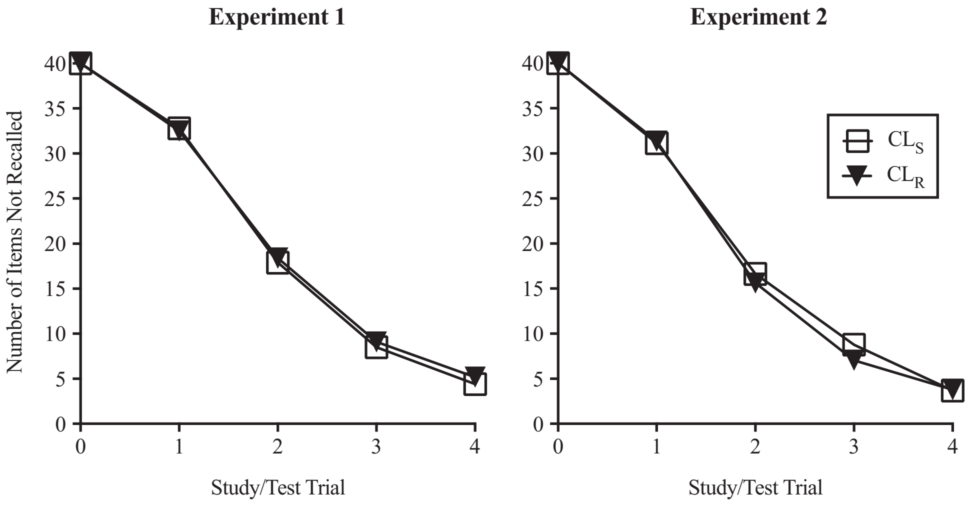

Figure 4. Average number of English word targets that had not been correctly recalled on any previous study/test trial during the Day 1 criterial learning task in Experiments 1 and 2. CL_S denotes participants who engaged in study practice after criterial learning, and CL_R denotes those who engaged in retrieval practice after criterial learning.

In addition to changing the retention interval, we also chose to manipulate stress induction within-subjects in Experiment 2. Because assessing memory during the first phase of the stress response was not a primary goal of this experiment, we opted to administer the two memory tests used in Experiment 1 in a different way in Experiment 2: pre- and post-stress. This allowed us to manipulate stress within-subjects; thus, decreasing error variance and improving the likelihood of detecting effects.

In Experiment 2, participants engaged in a criterial learning task followed by either additional restudying or retrieval practice. After a 1-week delay, they returned to the lab for stress induction with pre- and post-stress cued recall testing. In the context of the 1-week delay and the reduction of error because of our within-subjects manipulation of stress, we expected that additional retrieval practice would result in better performance on a final cued-recall test as compared with study practice, even in the presence of a delayed stress response.

Method

Design. The experiment used a 2 (learning strategy: CL_S or CL_R) × 2 (test timing: prestress or post-stress) mixed factorial design. Learning strategy was manipulated between subjects. Test condition was manipulated within subjects, such that all participants underwent stress induction. Nonstressed and stressed cued recall performance was measured by pre- and post-TSST-G tests.

Table 1
Average Cued Recall Performance on Test 1 and Test 2 in Experiment 1

Group	Test 1	Test 2
Nonstressed		
CL_S	16.0 (*.77*)	15.4 (*.65*)
CL_R	15.3 (*.92*)	15.1 (*.95*)
Stressed		
CL_S	15.3 (*.92*)	14.9 (*.92*)
CL_R	16.5 (*.62*)	15.3 (*.71*)

Note. *SE*s of the mean are in parentheses. CL_S denotes participants who engaged in study practice after criterial learning, and CL_R denotes those who engaged in retrieval practice after criterial learning.

Participants. Assuming an effect size of $\eta_p^2 = .04$ (Smith et al., 2016), a significance level of $\alpha = .05$, two between-subjects groups, two within-subjects measurements, and a conservative .70 correlation between repeated measures (in Experiment 1, the correlation between Test 1 and Test 2 performance was .90), an a priori power analysis recommended a total sample size of 50 participants to achieve 99% power to detect effects (G*Power 3.0; Faul et al., 2007). Sixty-five Tufts University students were initially recruited to participate in the experiment. However, seven participants were excluded from data analysis because they did not return for the second experimental session and two participants were excluded because they did not recall any items on one or both of the cued-recall tests on Day 2. Thus, all final analyses were conducted on 56 participants (35 women, $M_{\text{age}} = 19.54$, $SD_{\text{age}} = 1.38$). Participant recruitment and compensation was identical to Experiment 1. Twenty-seven participants were randomly assigned to the CL_S group and 29 were assigned to the CL_R group. All participants provided informed consent.

Materials and procedure. Testing sessions occurred on 2 days between 3:30 p.m. and 5:30 p.m. with a 1-week delay between Day 1 and Day 2 testing. All participants were tested two at a time. In a few cases in which only one participant showed up for the experiment, a research assistant served as a confederate in the experiment.

The Day 1 procedure (i.e., criterial learning and subsequent studying or retrieval practice) was identical to that of Experiment 1. One week later, participants returned to the original testing room where they completed a second STICSA and provided a baseline saliva sample. All participants were then given 2.5 min to complete Test 1, which was the same Test 1 used in Experiment 1. They then performed all tasks associated with TSST-G stress induction, which consisted of 2 min of speech preparation, 2 min each of speech delivery (4 min total), and 6 min of oral math subtraction. Participants then completed the third STICSA and provided the second saliva sample. They next viewed part of an episode of *The Office* during a 10-min break. Afterward, they provided the third saliva sample and were given 2.5 min to complete Test 2, which was the same Test 2 used in Experiment 1. Participants were then paid (when applicable), debriefed, and excused.

Cortisol measurement and data management. Saliva samples were stored and analyzed according to the procedures outlined in Experiment 1.

Results

Self-reported stress. As in Experiment 1, we first examined whether our Day 1 manipulation (CL_S or CL_R) affected participants' subsequent self-reported levels of stress. An independent samples *t* test on average Day 1 STICSA scores revealed no difference for participants who had engaged in additional retrieval practice versus additional study practice, $t(54) = 0.515$, $p = .609$.

To test whether the TSST-G tasks increased subjective anxiety on Day 2, we collapsed across learning group (CL_S or CL_R) and conducted a paired-samples *t* test comparing pre- and post-stress STICSA scores. As expected, participants demonstrated heightened post-stress scores relative to baseline, $t(55) = 2.665$, $p = .005$, $d = 0.37$. The average prestress STICSA score was 30.79 ($SEM = 1.20$) and the average post-stress score was 33.18 ($SEM = 1.42$).

Cortisol. To examine cortisol reactivity to the TSST-G, we collapsed across learning group (CL_S or CL_R) and conducted a one-way repeated measures ANOVA comparing the baseline cortisol measurement to that taken 12 and 30 min post-stress. As shown in Figure 5, we found a main effect of timing, $F(2, 106) = 12.467$, $p < .001$, $\eta_p^2 = .190$. Pairwise comparisons using a Bonferroni correction revealed significant cortisol increases from baseline to 12 min post-stress (mean difference = 0.752, $SEM = .212$, $p = .002$), from baseline to 30 min post-stress (mean difference = 1.712, $SEM = .456$, $p = .001$), and from 12 min post-stress to 30 min post-stress (mean difference = 0.960, $SEM = .319$, $p = .012$).

Day 1 criterial learning. As shown in Figure 4, there were no differences in the learning curves for participants who engaged in additional study practice or additional retrieval practice during the criterial learning task on Day 1. This was confirmed by a 2 (learning strategy: CL_S, CL_R) × 4 (study/Test Trial: 1, 2, 3, 4) mixed ANOVA with study/test trial as a within-subjects variable, in which there was no main effect of learning strategy on how many items participants correctly remembered on each cued-recall test, $F(1, 54) = 0.111$, $p = .740$. By the end of the task, participants had accurately recalled an average of 36.3 out of 40 items at least once, or approximately 91%.

Day 2 memory performance. For reasons discussed in Experiment 1, we included gender as a covariate in all of the following omnibus analyses. We conducted a 2 (learning strategy: CL_S or CL_R) × 2 (test timing: prestress or post-stress) ANCOVA, controlling for gender, to determine whether cued recall performance differed according to learning strategy or conditions of stress. We found a main effect of test timing, as participants recalled fewer words on the post-stress test ($M = 9.4$) than on the prestress test ($M = 7.9$), $F(1, 53) = 9.272$, $p = .004$, $\eta_p^2 = .149$. Replicating the robust testing effect, we also found a main effect of learning strategy, as those in the CL_R group recalled more words when averaged across the two tests ($M = 10.2$) than those in the CL_S group ($M = 7.1$), $F(1, 53) = 5.878$, $p = .019$, $\eta_p^2 = .100$. The interaction between learning strategy and test timing as well as the main effect of gender were nonsignificant (*F*s < 1). Average cued recall performance in Experiment 2 is displayed in Figure 6.

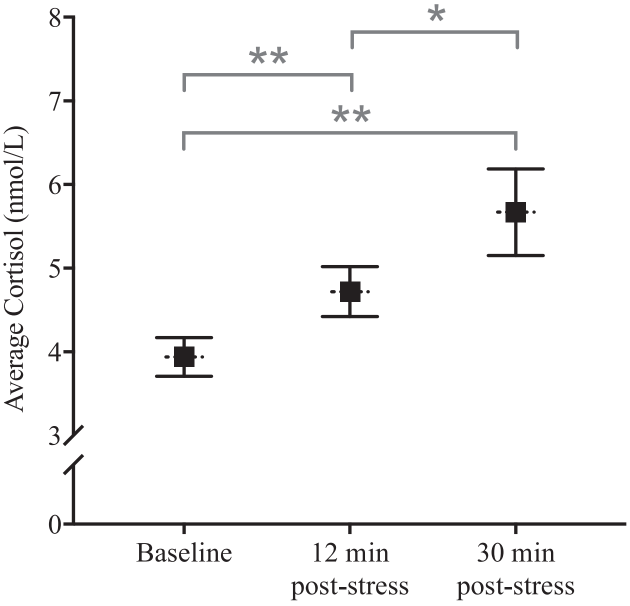

Figure 5. Average cortisol concentrations on Day 2 in Experiment 2. Cortisol samples were taken immediately before the TSST-G (Trier Social Stress Test for Groups; baseline), 12 min after the onset of the TSST-G, and 30 min after the onset of the TSST-G. Error bars represent *SEM*. $^{*}\ p < .05$. $^{**}\ p < .01$.

In summary, in Experiment 2, our stress manipulation again successfully induced subjective and physiological stress. In the context of a 1-week delay between encoding and retrieval, we found that additional retrieval practice after learning to criterion resulted in higher rates of recall than additional restudying. Further, regardless of the learning strategy used, participants demonstrated lower recall after stress than before stress.

General Discussion

The results of previous studies suggest that making multiple retrieval attempts when learning information can eliminate the detrimental impact that psychological stress typically has on memory retrieval (Schoofs & Wolf, 2009; Smith et al., 2016; Wolf et al., 2002). The primary goal of the present experiments was to determine whether many retrieval attempts are needed to buffer memory against stress, or whether one successful attempt is sufficient. When a 24-hr interval separated encoding and retrieval in Experiment 1, reaching a criterion of one correct recall attempt was equally as effective at supporting post-stress memory accessibility as three recall attempts. When that interval was expanded to a full week in Experiment 2, three recall attempts at encoding resulted in better post-stress performance than a single successful

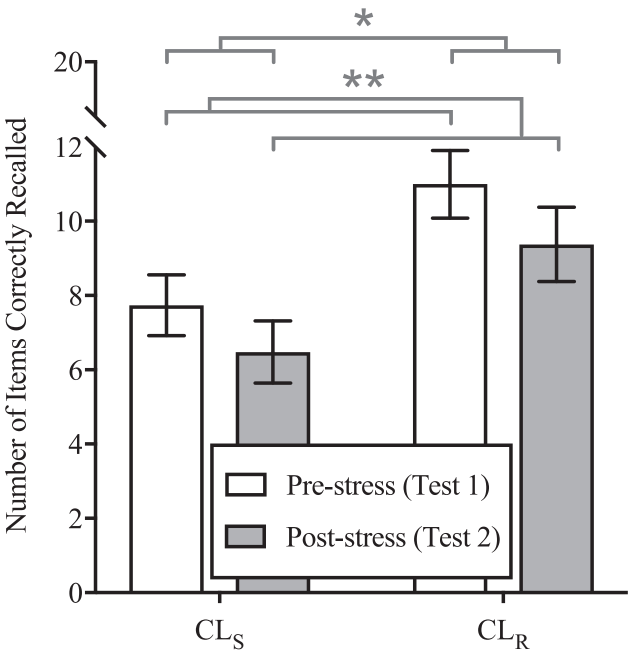

Figure 6. Average cued recall performance on Test 1 and Test 2 in Experiment 2. CL_S denotes participants who engaged in study practice after criterial learning, and CL_R denotes those who engaged in retrieval practice after criterial learning. Error bars represent *SEM*. $^{*}p < .05$. $^{**}p < .01$.

retrieval attempt. Thus, these results suggest that the number of retrieval attempts that are necessary to improve memory accessibility under stress is dependent on the amount of time that separates the retrieval attempts and the post-stress memory task.

The results of the present studies help further specify the relationship between retrieval practice as a learning strategy and memory retrieval in the context of stress. In Experiment 1, regardless of whether participants engaged in additional studying or retrieval practice after criterial learning on Day 1 or whether they were subjected to stress induction or a nonstressful task on Day 2, they demonstrated similar final cued recall performance. These findings stand in contrast to the majority of studies examining the effects of stress on memory, as stressed individuals did not demonstrate poorer memory performance than those who were not stressed. Our analysis of cortisol suggests that these discrepant findings are not because of an ineffective stress induction procedure; the majority of stressed participants demonstrated post-stress increases in cortisol and stressed participants as a group showed a strong post-stress increase in cortisol. Instead, these results suggest that, in the short-term, a single successful retrieval attempt during learning may be sufficient for creating stress-resistant memories. These results further refine the findings of Smith et al. (2016), who used a similar experimental paradigm but did not have participants learn to criterion before engaging in study practice or retrieval practice. Smith et al. (2016) found that three retrieval attempts during learning resulted in stress-resistant memories after a 24-hr delay, but simply restudying material during learning yielded the typical stress-related memory impairment. Our Experiment 1 findings suggest that, in the context of a 24-hr delay, learning to a criterion of one may provide the same memorial benefit as three retrieval attempts.

In Experiment 2, the retention interval between encoding and retrieval was increased from 24 hr to 1 week to determine whether criterial learning would continue to sufficiently support longer-term memory accessibility in the presence of stress. After implementing this 1-week delay, some of the hypotheses regarding the effects of the stress and criterial learning manipulations on episodic retrieval were borne out. Consistent with Karpicke and Roediger (2008), participants who engaged in additional retrieval practice after learning to criterion recalled more items than those who engaged in additional restudying. Further, replicating numerous studies (for reviews see Gagnon & Wagner, 2016; Shields et al., 2017), participants demonstrated lower recall after stress than before stress. However, in contrast to Smith et al. (2016), there was no interaction between learning strategy and test timing. Though retrieval practice improved memory relative to study practice, performance was negatively affected by stress in both groups. Thus, retrieval practice again emerged as the superior learning strategy for battling the negative effects of stress, but it did not completely buffer memory against the deleterious effects of stress as was observed by Smith et al. (2016).

The lack of an interaction between learning strategy and test timing may be explained by methodological differences between Experiment 1 and the study conducted by Smith et al. (2016). The first pertains to the stimuli used, as the participants in Smith et al. (2016) learned simple English nouns and images whereas those in our Experiment 1 learned Swahili-English word pairs. Foreign language learning requires the binding of cues with targets, a task for which we have less prior knowledge and fewer prior associations available than for memorizing words in our native language. Retrieval practice may serve to increase the ease with which information is recalled when stimuli are relatively easy to learn (i.e., Smith et al., 2016), but may not do so to the same degree when the stimuli require new associative infrastructure (i.e., Experiment 1). Researchers have debated the effectiveness of retrieval practice with increasingly complex stimuli (Karpicke & Aue, 2015; van Gog & Sweller, 2015), and we refer readers to van Gog and Sweller (2015) for a comprehensive review on the topic. Presently, under standard laboratory conditions, retrieval practice does appear to be efficacious with stimuli of varying complexities (e.g., Eglington & Kang, 2016; Martin, Nguyen, & McDaniel, 2016; Tempel, Loran, & Frings, 2015). The present results and those of Smith et al. (2016) may lend further insight into this debate. Specifically, stress may serve to expose memorial weaknesses that are not evident under nonstressful laboratory conditions. It is possible that retrieval practice creates better memory infrastructure for stimuli of low complexity (e.g., English nouns) than those of higher complexities (e.g., foreign language learning), but that this limitation is only evident when memories are subjected to stress. This hypothesis would benefit from a direct manipulation of stimulus complexity in the context of a paradigm such as that used in the present experiments.

An alternative but not mutually exclusive explanation for the lack of an interaction in Experiment 2 relates to the cognitive demand of the cued-recall test. Retrieving foreign vocabulary words is simply a more cognitively demanding task than retrieving individual words in one's native tongue. Further, post-stress memory performance has been shown to differ according to the level of cognitive demand that a test places on the test-taker. For example, tests of free recall tend to expose detrimental effects of stress more often than less cognitively demanding recognition tests (e.g., Buchanan & Tranel, 2008). Thus, the post-stress memory impairment observed in both learning groups in Experiment 2 may be attributable to our use of a memory test that was more demanding than that used by Smith et al. (2016). This explanation is well-supported by a wealth of studies demonstrating the impairing effects of stress on executive functions such as working memory and attention (for a review see Gagnon & Wagner, 2016). Because executive functions provide crucial support for retrieval processes (see Gagnon & Wagner, 2016; Levy & Anderson, 2002), more effortful retrieval tasks may be more vulnerable to stress. Because recalling Swahili-English word pairs may be more difficult than recalling English nouns, this retrieval task may be more vulnerable to stress regardless of how information was learned. To further test this hypothesis, future researchers may consider manipulating cognitive load during retrieval in a stress-and-memory paradigm.

Finally, our within-subjects manipulation of stress in Experiment 2 may have also contributed to our null interaction between test timing and learning strategy. Relative to between-subjects designs, like that used by Smith et al. (2016), within-subjects manipulations reduce error variance and better control for individual differences in performance. Thus, if stress does indeed impair memory retrieval for individuals who engaged in retrieval practice at encoding such an effect may be better detected using a within-subjects experimental design such as that used in Experiment 2.

In addition to the theoretical contributions of the present experiments, our results begin to shed light on the practical implications of using retrieval practice to support post-stress memory. As mentioned, college students who use self-testing to study for exams often stop studying once they have learned to a criterion of one (Kornell & Bjork, 2007, 2008; Wissman et al., 2012). The results of Experiment 1 suggest that this approach may provide sufficient memory support for students cramming for a stressful next-day exam. However, for those studying for exams that are several days or weeks away, such as high-stakes standardized tests, the results of Experiment 2 suggest that additional retrieval practice may yield the best memory access. This advice is consistent with the broader retrieval practice literature, in which learning to a criterion of three (i.e., three successful retrieval attempts) is recommended for optimizing long-term memory and time spent studying (Rawson & Dunlosky, 2011).

References

Beckner, V. E., Tucker, D. M., Delville, Y., & Mohr, D. C. (2006). Stress facilitates consolidation of verbal memory for a film but does not affect retrieval. *Behavioral Neuroscience, 120,* 518–527. http://dx.doi.org/10.1037/0735-7044.120.3.518

Buchanan, T. W., & Tranel, D. (2008). Stress and emotional memory retrieval: Effects of sex and cortisol response. *Neurobiology of Learning and Memory, 89,* 134–141. http://dx.doi.org/10.1016/j.nlm.2007.07.003

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5,* 781. http://dx.doi.org/10.3389/fpsyg.2014.00781

Eglington, L. G., & Kang, S. H. K. (2016). Retrieval practice benefits deductive inference. *Educational Psychology Review, 30,* 215–228.

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

Gagnon, S. A., & Wagner, A. D. (2016). Acute stress and episodic memory retrieval: Neurobiological mechanisms and behavioral consequences. *Annals of the New York Academy of Sciences, 1369,* 55–75. http://dx.doi.org/10.1111/nyas.12996

Gärtner, M., Rohde-Liebenau, L., Grimm, S., & Bajbouj, M. (2014). Working memory-related frontal theta activity is decreased under acute stress. *Psychoneuroendocrinology, 43,* 105–113. http://dx.doi.org/10.1016/j.psyneuen.2014.02.009

Grös, D. F., Antony, M. M., Simms, L. J., & McCabe, R. E. (2007). Psychometric properties of the state-trait inventory for cognitive and somatic anxiety (STICSA): Comparison to the state-trait anxiety inventory (STAI). *Psychological Assessment, 19,* 369–381. http://dx.doi.org/10.1037/1040-3590.19.4.369

Herlitz, A., Nilsson, L. G., & Bäckman, L. (1997). Gender differences in episodic memory. *Memory & Cognition, 25,* 801–811. http://dx.doi.org/10.3758/BF03211324

Hupbach, A., & Fieman, R. (2012). Moderate stress enhances immediate and delayed retrieval of educationally relevant material in healthy young men. *Behavioral Neuroscience, 126,* 819–825. http://dx.doi.org/10.1037/a0030489

JASP Team. (2017). JASP (Version 0.8.2). [Computer software]. Amsterdam, the Netherlands: Author.

Jeffreys, H. (1961). *The theory of probability* (1st/3rd ed.). Oxford, England: Clarendon Press. (Orginal work published 1939)

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review, 27,* 317–326. http://dx.doi.org/10.1007/s10648-015-9309-3

Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science, 319,* 966–968. http://dx.doi.org/10.1126/science.1152408

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14,* 219–224. http://dx.doi.org/10.3758/BF03194055

Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: The benefits - and costs - of dropping flashcards. *Memory, 16,* 125–136. http://dx.doi.org/10.1080/09658210701763899

Levy, B. J., & Anderson, M. C. (2002). Inhibitory processes and the control of memory retrieval. *Trends in Cognitive Sciences, 6,* 299–305. http://dx.doi.org/10.1016/S1364-6613(02)01923-X

Lewin, C., Wolgers, G., & Herlitz, A. (2001). Sex differences favoring women in verbal but not in visuospatial episodic memory. *Neuropsychology, 15,* 165–173. http://dx.doi.org/10.1037/0894-4105.15.2.165

Lovallo, W. R., Robinson, J. L., Glahn, D. C., & Fox, P. T. (2010). Acute effects of hydrocortisone on the human brain: An fMRI study. *Psychoneuroendocrinology, 35,* 15–20. http://dx.doi.org/10.1016/j.psyneuen.2009.09.010

Martin, N. D., Nguyen, K., & McDaniel, M. A. (2016). Structure building differences influence learning from educational text: Effects on encoding, retention, and metacognitive control. *Contemporary Educational Psychology, 46,* 52–60. http://dx.doi.org/10.1016/j.cedpsych.2016.03.005

Oei, N. Y., Everaerd, W. T., Elzinga, B. M., van Well, S., & Bermond, B. (2006). Psychosocial stress impairs working memory at high loads: An association with cortisol levels and memory retrieval. *Stress, 9,* 133–141. http://dx.doi.org/10.1080/10253890600965773

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60,* 437–447. http://dx.doi.org/10.1016/j.jml.2009.01.004

Qin, S., Hermans, E. J., van Marle, H. J., Luo, J., & Fernández, G. (2009). Acute psychological stress reduces working memory-related activity in the dorsolateral prefrontal cortex. *Biological Psychiatry, 66,* 25–32. http://dx.doi.org/10.1016/j.biopsych.2009.03.006

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General, 140,* 283–302. http://dx.doi.org/10.1037/a0023956

Reul, J. M., & de Kloet, E. R. (1985). Two receptor systems for corticosterone in rat brain: Microdistribution and differential occupation. *Endocrinology, 117,* 2505–2511. http://dx.doi.org/10.1210/endo-117-6-2505

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15,* 20–27. http://dx.doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181–210. http://dx.doi.org/10.1111/j.1745-6916.2006.00012.x

Salimetrics, LLC. (2017). Retrieved from www.salimetrics.com

Schneider, W., Eschman, A., & Zuccolotto, A. (2001). *E-Prime user's guide*. Pittsburgh, PA: Psychology Software Tools.

Schönfeld, P., Ackermann, K., & Schwabe, L. (2014). Remembering under stress: Different roles of autonomic arousal and glucocorticoids in memory retrieval. *Psychoneuroendocrinology, 39,* 249–256. http://dx.doi.org/10.1016/j.psyneuen.2013.09.020

Schoofs, D., & Wolf, O. T. (2009). Stress and memory retrieval in women: No strong impairing effect during the luteal phase. *Behavioral Neuroscience, 123,* 547–554. http://dx.doi.org/10.1037/a0015625

Schwabe, L., Joëls, M., Roozendaal, B., Wolf, O. T., & Oitzl, M. S. (2012). Stress effects on memory: An update and integration. *Neuroscience and Biobehavioral Reviews, 36,* 1740–1749. http://dx.doi.org/10.1016/j.neubiorev.2011.07.002

Schwabe, L., & Wolf, O. T. (2014). Timing matters: Temporal dynamics of stress effects on memory retrieval. *Cognitive, Affective & Behavioral Neuroscience, 14,* 1041–1048. http://dx.doi.org/10.3758/s13415-014-0256-0

Shields, G. S., Sazma, M. A., McCullough, A. M., & Yonelinas, A. P. (2017). The effects of acute stress on episodic memory: A meta-analysis and integrative review. *Psychological Bulletin, 143,* 636–675. http://dx.doi.org/10.1037/bul0000100

Smith, A. M., Floerke, V. A., & Thomas, A. K. (2016). Retrieval practice protects memory against acute stress. *Science, 354,* 1046–1048. http://dx.doi.org/10.1126/science.aah5067

Tempel, T., Loran, I., & Frings, C. (2015). Dancing your moves away: How memory retrieval shapes complex motor action. *Journal of Experimental Psychology: Applied, 21,* 300–312. http://dx.doi.org/10.1037/xap0000052

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. http://dx.doi.org/10.1037/0278-7393.4.3.210

van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review, 27,* 247–264. http://dx.doi.org/10.1007/s10648-015-9310-x

von Dawans, B., Kirschbaum, C., & Heinrichs, M. (2011). The Trier Social Stress Test for Groups (TSST-G): A new research tool for controlled simultaneous social stress exposure in a group format. *Psychoneuroendocrinology, 36,* 514–522. http://dx.doi.org/10.1016/j.psyneuen.2010.08.004

Weitzman, E. D., Fukushima, D., Nogeire, C., Roffwarg, H., Gallagher, T. F., & Hellman, L. (1971). Twenty-four hour pattern of the episodic secretion of cortisol in normal subjects. *The Journal of Clinical Endocrinology and Metabolism, 33,* 14–22. http://dx.doi.org/10.1210/jcem-33-1-14

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory, 11,* 571–580. http://dx.doi.org/10.1080/09658210244000414

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory, 20,* 568–579. http://dx.doi.org/10.1080/09658211.2012.687052

Wolf, O. T., Schommer, N. C., Hellhammer, D. H., Reischies, F. M., & Kirschbaum, C. (2002). Moderate psychosocial stress appears not to impair recall of words learned 4 weeks prior to stress exposure. *Stress, 5,* 59–64. http://dx.doi.org/10.1080/102538902900012332

Received January 5, 2018
Revision received February 21, 2018
Accepted February 22, 2018 ■