Setting the alarm: Word emotional attributes require consolidation

to be operational

Nicolas Dumay[1,2], Dinkar Sharma[3,] Nora Kellen[3], and Sarah Abdelrahim[3]

[1]*Department of Psychology, University of Exeter, United Kingdom;*

[2]*BCBL. Basque Center on Cognition, Brain and Language, Spain;*

[3]*School of Psychology and Centre for Cognitive Neuroscience and Cognitive Systems, University of Kent, United Kingdom*

Correspondence:

Nicolas Dumay

Department of Psychology,

University of Exeter,

Perry Road,

EX4 4QG

Exeter

United Kingdom

E-mail: nicolas.dumay@gmail.com

Phone: +44 (0) 1392 724666

Fax: +44 (0) 1392 724623

Abstract

Demonstrations of emotional Stroop in conditioned made-up words are flawed due to the lack of task ensuring similar word encoding across conditions. Here, participants were trained on associations between made-up words (e.g., 'drott') and pictures with an alarming or neutral content (e.g., 'a dead sheep' versus 'a munching cow') in a situation that required attention to both ends of each association. To test whether word emotional attributes need to consolidate before they can hijack attention, one set of associations was learnt seven days before the test, whereas the other set was learnt either six hours or immediately before the test. The novel words' ability to evoke their emotional attributes was assessed using both Stroop and an auditory analogue called pause detection. Matching words and pictures was harder for alarming associations. However, similar learning rate and forgetting at seven days were observed for both types of associations. Pause detection revealed no emotion effect for same-day (i.e., unconsolidated) associations, but robust interference for seven-day-old (i.e., consolidated) alarming associations. Attention capture was found in the emotional Stroop as well, though only when trial *n-1* referred to a same-day association. This task also showed stronger response repetition priming (independently of emotion) when trials *n* and *n-1* both tapped into seven-day-old associations. Word emotional attributes hence take between six hours and seven days to be operational. Moreover, age interactions between consecutive trials can be used to gauge implicitly the indirect (relational) episodic associations that develop in the meantime between the memories of individual items.

Keywords: emotional Stroop; memory consolidation; threat detection; relational memory; response/task conflict

1. Introduction

Assessing the ease with which a person reports the colour in which particular types of words are printed has been a useful diagnostic tool in clinical populations (for reviews, see Bar-Haim, Lamy, Pergamin, Bakermans-Kranenburg, & Van Ijzendoorn, 2007; Williams, Mathews, & MacLeod, 1996; Yiend, 2010). For example, depressed individuals have a much harder time than non-depressed individuals naming or classifying (by button push) the colour print of words associated with depression (e.g., defeat, failure, reject), whereas post-traumatic stress disorder patients (e.g., rape victims, veterans, etc.) have, instead, more difficulties with words associated with the specific trauma that they have experienced (e.g., Caparos & Blanchette, 2014; Epp, Dobson, Dozois, & Frewen, 2012; Khanna et al., 2015; Pergamin-Hight, Naim, Bakermans-Kranenburg, van Ijsendoorn, & Bar-Haim, 2015; for reviews, see Cisler et al., 2011; Joormann, 2010). A similar effect is observed for specific word in those suffering from panic attacks, obsessive-compulsive behaviours, generalized anxiety, addiction, and phobias (e.g., Cox, Fadardi, & Pothos, 2006; Phaf & Kan, 2007; Williams, Watts, MacLeod & Mathews, 1997). It is also obtained from threat-related words (e.g., death) in nonpathological individuals, suggesting a predisposition to react to threat by a reallocation of attention (for a review, see Cisler & Koster, 2010). It can surface not only as a capture of attention on the current trial (a fast component; e.g., Pratto & John, 1991), but also—and sometimes more prominently—as a difficulty in disengaging attention from the word's emotional content accessed on the previous trial (a slow component; e.g., McKenna & Sharma, 2004; see also Bertels & Kolinsky, 2015). This affective interference has been referred to as the 'emotional Stroop', due to its similarity, in appearance at least, with the classic Stroop effect (cf. Dalgleish, 2005).

However, unlike the classic Stroop (Stroop, 1935; MacLeod, 1991), which is measured at the item-level, by presenting the same colour word (e.g., red) in a colour-congruent (red) and a colour-incongruent condition (blue), the emotional Stroop is almost unavoidably measured by comparing performance between two different sets of items: the emotional words and the neutral words. This raises the possibility that emotional Stroop has nothing to do with emotion, and is, instead, driven by some systematic uncontrolled difference(s) between the orthographic, phonological and/or lexico-semantic properties of the two word sets (cf. Bowers, Davis, & Hanley, 2005; Cutler, 1981;

Magnuson, Dahan, Tanenhaus, & Aslin, 2003, for earlier warnings). Thus, any perceptual or linguistic dimension that affects how the word is processed and covaries with emotional content could drive the effect.

Evidence showing that this is clearly a possibility comes from studies looking at unintentional word reading, by means of non-colour word Stroop tasks (for a review, see Kinoshita, De Wit, & Norris, 2017). These studies demonstrate that the extent to which the orthographic make-up of a word (i.e., its sublexical letter sequences) invites reading aloud is a prime determinant of colour classification latencies: these are slower for letter strings that are most prototypical of the language (e.g., Klein, 1964; Monsell, Taylor, & Murphy, 2001). As Kinoshita et al. also showed, reading the word (and thus having a task conflict) is more likely to occur when the modality of colour response also involves phonology (i.e., vocal responses); as this interference is driven by pronounceability, it increases the more there are letters in the word. In addition, frequency manipulations (e.g., Burt, 2002; Monsell et al., 2001) have shown a (not always reliable) trend, whereby words seen often in the language are colour-named faster than words seen less often. Thus, subtle (uncontrolled) differences in the linguistic make-up of the words could be behind the emotional Stroop.

The bomb was dropped by Larsen, Mercer, and Balota (2006), whom, in a meta-analysis, uncovered some worrying confounds in the lexical properties and processing of the negative, positive, disorder-specific and neutral (control) words of 32 published emotional Stroop studies (i.e., 1,033 unique words in total). They found that negative and disorder-specific words were less frequent than neutral words, and that emotional words, whether negative or positive, also had more letters in them. Relying on Balota et al.'s (2002) corpus of behavioural data, the authors found that negative and disorder-specific words returned poorer lexical decision and naming performance compared to neutral words (see their Table 1). Based on the non-colour word Stroop findings reviewed above, this is exactly what one would expect. Although one could argue that these behavioural differences may themselves reflect the influence of emotions, they actually go in the opposite direction to what an attention-grabbing mechanism would produce. Because, in tasks such as naming and lexical decision, threat directs attention straight onto the relevant dimension (i.e., the word), better performance for negative and disorder-specific words should be observed, not the reverse. Except for disorder-specific

words, Larsen et al. showed that the behavioural differences were entirely explained by the lexical properties of the items.

The Larsen et al. (2006) study casts doubt on the emotional nature of the emotional Stroop. It suggests instead that the observed interference reflects systematic differences in the processing requirements between emotional and nonemotional words. One counter-argument, however, is that in clinical populations the interference produced by disorder-specific words returns to baseline as the symptoms disappear. Although this is compatible with an emotion-based interpretation, one could equally imagine that, with a pathology emerging, the content of one person's language would include more exposure to words relating to the area of concern, and this could be picked up by the Stroop test. That the interference goes away while patients are still exposed more than usual to some of these words during therapy could be a way out for the emotional hypothesis. However, this would require demonstration that the amount of exposure received in therapy session could make up for the fact that at this point the words in question no longer occur as frequently in the patient's language.

To tackle the issue, Richards and Blanchette (2004) manipulated emotionality within item by imposing new emotional attributes to a set of existing (neutral) words (e.g., 'camera') and nonwords (e.g., 'patarel') via associative learning.[1] Across trials, words and pseudowords always appeared together with either emotional or neutral pictures, so that ultimately the linguistic stimuli could inherit the emotional value of the picture set they were paired with. All negative associations were learnt first, or vice versa. In the Stroop task that followed, negatively conditioned pseudowords produced interference only in the most anxious participants, whereas words produced no effect whatsoever.

These results are important because they suggest that with just a handful of exposures (there were actually five per word), a letter string at first unknown to the brain can acquire emotional attributes and by itself (i.e., in the absence of the unconditioned image) trigger a threat reaction. They support the idea that, despite the presence of lexical (and potentially other linguistic) confounds, emotional

[1] We cannot help but note that out of the twelve nonwords used by Richards and Blanchette (2004), eight were close orthographic (substitution) neighbours of existing words (i.e., gruss, fronded, partled, tounded, admo, trovoke, broubled, donger), six of which had a negative valence (i.e., gross, paroled, wounded, provoke, troubled, danger). These lexical contaminations would make it hard to observe an associative emotional Stroop effect within these items, as it would mean for the neutral condition to go against the natural negative valence infused by the existing neighbour.

Stroop reflects the capture of attention by emotions. Hence, Richards and Blanchette concluded that "… emotional connotations are a vital component of the emotional Stroop effect" (pp. 279-280).

These data were reproduced in cocaine users by Sharma and Money (2010), using the same learning procedure (except that valence was not blocked). In a variant of the emotional Stroop administered soon after, in which every critical trial was followed by six neutral fillers (McKenna & Sharma, 2004), they found that nonwords conditioned with addiction images interfered with colour naming, not on the current trial but on the subsequent trial. As the authors suggested, craving may have been responsible for some difficulties in disengaging attention from drug-conditioned stimuli; alternatively, drug-conditioned stimuli may have reactivated the specific anxiety associated with the outlaw nature of the addiction. In either case, the underlying assumption is that attention is hijacked because of the word reactivating the associated emotions (respectively, pleasure and threat).

In a recent replication of their own findings, Blanchette and Richards (2013) reported puzzling elements, however. While they repeatedly observed threat-conditioned Stroop interference in anxious participants, ratings of the affective valence of the nonwords immediately after the Stroop test showed no difference between conditions, whether overall or just in the most anxious participants. This is puzzling because one would expect the effects of emotion conditioning to be easier to capture via explicit judgements than via a task where emotions are irrelevant, like colour naming. Consequently, one may wonder whether the interference obtained through this conditioning procedure really is driven by the evocation of freshly-learnt emotional attributes. Blanchette and Richards interpreted the above dissociation in terms of multiple (independent) levels of reaction to threat. Because expressive and physiological markers of emotions (i.e., face muscle contraction and skin conductance) had higher values for negative than neutral nonwords during the Stroop task and across the whole sample, the story seems to fit. However, a closer look at the protocol suggests another explanation, with at least two possible scenarios.

In the Richards and Blanchette (2004) procedure used in the above studies, participants saw a picture for two seconds, with a word (or a nonword) appearing after 500 ms, superimposed on it centrally until the end of the trial (see Fig. 1a). Participants were only instructed to pay close attention to both the picture and the word until they disappeared. Hence, experimenters had little control over

how much encoding was done of each dimension. Therefore, the possibility remains that the balance between looking and paying attention to the word versus the picture varied as a function of emotionality. For instance, in the presence of a negative picture, participants could have had a hard time encoding the word because of the attention-grabbing nature of the background image. Bisby and Burgess (2014), for example, showed that if negative pictures have a long-term memory advantage over neutral pictures, this is at the expense of the associated context: 24 hours after encoding, their participants had better memory of the negative target pictures, but poorer memory of the neutral background. Given the potential for a similar 'weapon focus' effect during encoding in the Richards and Blanchette study (2004; e.g., Easterbrook, 1959; Kensinger, 2009), nonwords paired with negative pictures could have looked less familiar in the emotional Stroop test that followed. This could have caused surprise, leading to longer colour classification latencies.[2]

Conversely, because, during conditioning, negative and neutral pictures always appeared in separate blocks and no task was required, participants could have adapted to the emotional context and learnt to focus on the linguistic dimension when the background was unpleasant. As a result, in the emotional Stroop test, letter strings encoded in the presence of a negative picture could have produced more resonance from episodic memory, again dragging participants away from the task.[3]

This view of emotion modulating how much of the letter string is encoded correctly predicts that the Stroop interference should be more visible on nonwords than on words. As nonwords are, by definition, new to memory, they would provide more leg room to index possible variations in strength of encoding. That the interference is seen only in the most anxious participants is also expected: if words are less well encoded in the negative condition because the background picture is grabbing attention, then this effect should be the largest in high-anxiety individuals, as these are precisely the ones who would be the most disturbed by the presence of an alarming distractor. In selective attention tasks with spatially distinct emotional (or just salient) distractors, nonpathologically anxious individuals show enhanced amygdala reactivity, coupled with poorer recruitment of emotion/cognitive

[2] We thank one of the reviewers for this suggestion.
[3] The fact that in Sharma and Money (2010) the Stroop interference was obtained on the subsequent (instead of on the current) trial does not make any of these explanations obsolete.

control areas, compared to nonanxious individuals (e.g., Bishop, Duncan, & Lawrence 2004; Bishop, Duncan, Brett, & Lawrence, 2004; for reviews, see Eysenck, Derakshan, Santos, & Calvo, 2007; Öhman, 2005; Carretié, 2014). If the effect reflects cognitive avoidance, it should also be most visible in high-anxiety participants. Despite a possibly reduced engagement of their emotion control areas (e.g., Bishop, Duncan, Brett et al., 2004), these individuals would be the ones most likely to focus on the word in order to avoid the unpleasantness of picture. In Blanchette and Richards (2013), emotion-driven changes in facial expression *during conditioning* were found indeed only in the most anxious participants. In short, whatever the scenario, our view of the above conditioning data assumes a genuine effect of emotion, but one that takes place during encoding (i.e., in the presence of the emotional dimension), not during the colour classification Stroop test.[4]

What is also surprising about the above results is that they seem to indicate that word emotional attributes are operational straight away, without the need for an offline consolidation interval (for a review, see Wixted & Cai, 2014). Word learning studies, however, show that memory consolidation, in particular during sleep, is pivotal in linking novel word forms with semantic information. In particular, this is seen in the new words being unable to activate their meaning automatically before the next day (i.e., once sleep has occurred). For instance, Clay, Bowers, Davis, and Hanley (2007) showed that semantic interference from novel word distractors in naming objects (e.g., a strawberry) is not observed immediately after learning the made-up words and their meaning (e.g., a 'kosla' is a bitter and spiky fruit), but only after a delay. (In their experiment, re-test occurred after a week.) This result was recently corroborated by Geukes, Gaskell, and Zwitserlood (2015; Experiment 3) using the colour-Stroop task. Novel words learnt in association with existing colour words (e.g., 'alep-blau'; this was done in German) were found to produce interference in colour naming (e.g., 'alep' printed in red

[4] Sharma and Money (2010) trained their participants until they reached 100 percent correct at a memory test assessing their knowledge of the associations. While at first this may suggest equal encoding of letter strings in the two conditions, this is no guarantee. As the inclusion criteria was at ceiling, the authors had no means of telling whether one condition was better learnt than the other. Further, as participants were required only to report which category each nonword belonged to, this task did not give any information about the level of acquisition of each association taken individually. In fact, it could be performed just on the basis of familiarity with the nonword or the extent to which the picture can be remembered, both of which are likely to depend on the emotional content of the image.

vs. blue), but this was only for associations learnt on the previous day, not for associations learnt on the same day as the test. Polysomnographic recordings by Tamminen, Lambon Ralph, and Lewis (2013) have implicated sleep (via changes in spindle activity[5]) as the driving force behind the entrenchment of novel lexical concepts in semantic memory. As emotional attributes, like semantic attributes, are necessarily abstracted from episodic knowledge, these findings make it unlikely to find language-mediated threat reactivation (i.e., in the absence of episodic cues) immediately after encoding, as in Richards and Blanchette (2004), for instance.

But the literature on memory consolidation allows us to make another projection. Sleep is known to specifically promote consolidation of emotionally salient memories (for reviews, see Goldstein & Walker, 2014; Payne & Kensinger, 2010). For example, Wagner, Gais and Born (2001) found that specifically REM-rich late-night sleep helped retention of negative narratives compared to neutral narratives. Similarly, Hu, Stylos-Allan, and Walker (2006) showed that negative photographs looked more familiar after 12 hours with sleep than after an equivalent interval spent awake, an effect which neutral photographs learnt concomitantly did not show (see Nishida, Pearsall, Buckner, & Walker (2009) for REM-sleep oscillatory correlates of this effect). Importantly, as demonstrated by Payne, Stickgold, Swanberg, and Kensinger (2008), sleep modifies emotional memories by enhancing the negative aspects of the scene (e.g., a damaged car) at the expense of its neutral aspects (e.g., pedestrians on the pavement; see also Cunningham et al., 2014; Payne, Chambers, & Kensinger, 2012; Payne et al., 2015). In view of this overnight memory trade-off, a straightforward prediction is that negative memories should be even more negative once sleep has occurred. Hence, if consolidation plays, emotion effects of episodic memories reactivated by presenting an associated cue, such as a word, should be bigger by the next morning.

2. The present study

The present study hence re-assessed whether the emotional attributes of new words are operational straight away, or whether instead they require consolidation before they can hijack attention. To test for consolidation, participants learnt Sets 1 and 2 of novel-word/picture associations one week apart

[5] Spindles are bursts of oscillatory brain activity visible on the EEG and typical of Stage 2 sleep.

and were tested on both sets shortly after learning Set 2 on that same day (see Fig. 2 for a sketch of the protocol). Thus, Set 1 was seven-day old and had plenty of time to consolidate by the time the test occurred, whereas Set 2 had just been learnt. This ensured a direct assessment of consolidation/forgetting uncontaminated by the effects of test practice (see Takashima et al., 2009). Within each set, half of the associations were emotionally negative and the other half were neutral. In contrast to Richards and Blanchette (2004), negative and neutral trials were all intermixed, and participants had to learn which picture amongst several of the same emotional category referred to the novel word presented. This ensured that participants paid attention to both dimensions on each trial (see Fig. 1b for an example trial with three distractors). After the response, the correct picture remained on the screen, so that participants could learn the association. Feedback was turned off on the last round of trials to measure final acquisition levels.

To boost learning, novel words were presented both visually and auditorily, so that the two modalities reinforced each other. This also meant that effects of attentional capture could be tested in both modalities. Consequently, at test, the colour Stroop task was followed by an auditory equivalent, called 'pause detection', in which participants had to decide whether a short artificial disruption was present within the audio file. As Mattys and Clark (2002) showed, latencies at detecting these artificial silences are positively correlated with how much attentional resources are captured by comprehending speech. Therefore, we expected this task also to index hijacking of attention due to threat reactivation. The test phase ended with one more round of novel-word/picture association to re-assess explicit knowledge.

Finally, exposing participants to alarming/disturbing pictures may have a number of short-term consequences at multiple levels. These include increased alertness and sensitivity to negative and/or unfamiliar stimuli, heightened threat-related expectations, a shift in processing mode with a stronger weight on emotionality, and depleted (or, on the contrary, more efficient) decision making. To explore this possible 'sensitization', half of the participants (i.e., the 0-hr group) were tested immediately after learning Set 2, whereas the other half (i.e., the 6-hr group) were tested six hours later.

If the emotional Stroop effect is emotional in nature, latencies in the colour classification and the pause detection task should be slower for negative (compared to neutral) novel words. In addition, if

word emotional attributes require consolidation to be capable of hijacking attention, then the interference should not be seen on same-day associations, but only on seven-day-old associations. Finally, if exposure to the negative pictures of Set 2 make participants momentarily more fragile and/or more alert and tuned to emotions, then larger emotion effects (especially those triggered by consolidated Set 1) may be found for the 0-hr group, tested straight after.

3. Method

3.1 Participants

Sixty-four UK-English native speakers (21 males; age range: 17-34), all of whom were students at the University of Kent, were tested. Half of them were assigned to the 0-hr group, and the other half to the 6-hr group. All of them reported not to be suffering from any psychiatric condition, nor to be under medication likely to alter their mood. They also had no known auditory, language or sleep impairment, and their eye-sight was normal or corrected-to-normal. All reported to have slept at least six hours the nights preceding Day 1 and Day 8, and the night following Day 1, and to have gone to bed by 23:00 on these evenings. All participants were paid for taking part, at £7/hr. Informed consent was obtained from all in a manner approved by the University of Kent Ethics Committee.

3.2 Materials

The 40 picture stimuli came from the International Affective Picture System (henceforth, IAPS; Lang, Bradley, & Cuthbert, 2008; see Appendix A) and were split into two sets. Within each set, half the pictures had a low (i.e., negative) valence and a high level of arousal, whereas the other half had a mid-range (i.e., neutral) valence and a low level of arousal. Pictures that had a sexual connotation were avoided. According to the IAPS ratings, our two negative subsets were matched on valence (2.17 vs. 2.18; $t(18) = .08$, $p > .93$) and arousal (6.06 vs. 5.97; $t(18) = .34$, $p > .73$), and so were our two neutral subsets (5.20 vs. 5.31; $t(18) = .56$, $p > .58$; 3.53 vs. 3.95; $t(18) = .97$, $p > .34$), and within each set negative and neutral pictures differed significantly on the two dimensions (valence: $ts(18) > 17.60$, $ps < .0001$; arousal: $ts(18) > 4.75$, $ps < .0002$).

To confirm that our four subsets were matched (vs. highly contrastive) on both valence and arousal, 40 participants (18-27 years old; 25 females) from the same population as the main sample (see Section 3.1 for inclusion criteria) rated the stimuli. Participants were shown each picture for three

seconds, followed by two 9-point Likert scales, one for valence (1 as the most negative; 9 as the most positive) and the other for arousal (1 as the most calming; 9 as the most arousing), which they used in a self-paced fashion. Trial order was randomized for each subject. Based on these ratings also, the two negative subsets were matched (valence: 2.07 vs. 2.12; $t(18) = .17$, $p > .86$; arousal: 7.34 vs. 7.24; $t(18) = .34$, $p > .73$), and so were the two neutral subsets (5.66 vs. 5.33; $t(18) = 1.15$, $p > .26$; 4.44 vs. 4.66; $t(18) = .70$, $p > .49$), and within each set the negative and neutral pictures differed on the two dimensions (valence: $ts(18) > 10.38$, $ps < .0001$; arousal: $ts(18) > 6.80$, $ps < .0001$).[6] In sum, our four subsets of IAPS pictures were tailored to our student population. Means for valence and arousal for each picture as per these ratings are reported in Appendix A.

The novel words were 80 nonsense monosyllables with a length of five or six letters and of three or four phonemes (e.g., 'drott'; see Appendix A). They were all orthographic hermits (i.e., from which no real word can be created by adding, subtracting or substituting one letter), both visually and auditorily plausible in English. The rationale for using hermits was that we did not want our novel words to inherit the semantic and emotional attributes of an English word (see Camblats & Mathey, 2016); for this reason we further checked that none of them bore any resemblance overall to an English word. The spoken stimuli were digitally recorded (16-bit/44.1kHz) in a soundproof booth by a female speaker, using a Sennheiser ME65 microphone and a Tascam HD-P2.

3.3 Design

Out of the 80 novel words, 40 were presented during the first or the second exposure session in association with one of the 40 pictures; the other words were presented only during the Stroop and pause detection tests together with the trained items and were used as a baseline. To ensure that every item contributed equally to all conditions, they were split into eight groups matched mechanically (all $Fs < 1$; van Casteren & Davis, 2007) on length in letters (mean: 5.7 (0.1)) and phonemes (mean: 3.8 (0.0)), phonological neighbourhood (mean: 6.4 (0.2)) and bigram positional token frequency (mean: 18,962.2 (933.2)) given by N-Watch (Davis, 2005). The four picture subsets were rotated 'clockwise'

[6] The strong correlation coefficients ($r = .97$ for valence; $r = .89$ for arousal) between these and the IAPS ratings suggest that our more positive values for the neutral pictures and our higher arousal values for the negative pictures reflect procedural differences rather than genuine discrepancies in the emotionality of these pictures between the two populations.

over these eight item groups across eight versions of the experiment. For half the participants (in both the 0-hr and the 6-hr group), one set of 20 pictures (i.e., ten negative and ten neutral) and associated words was learnt seven days before the test, whereas the other set was learnt on the same day as the test, and the reverse applied to the other half of the participants. To counterbalance pause presence/absence in the pause detection test, the eight item groups were further split into two, leading to 16 versions of the experiment.

3.4 Procedure

The experiment was run any time between 9:00 and 20:00. However, to avoid circadian effect on encoding, for any given participant exposure to Set 2 occurred at the same time of the day as exposure to Set 1. Time stamps from the experimental software showed that on average exposure to Set 1 was completed by 12:09 and exposure to Set 2 was completed by 12:11 ($t(63) = .67$, $p > .50$). As participants in the 6-hr group were tested six hours after their exposure to Set 2, encoding for these participants was completed on average two hours earlier (11:19 (range: 09:25-13:01) compared to the 0-hr group (13:09 (09:25-17:51), $t(62) = 3.90$, $p < .001$). That also means that their test phase occurred four hours later than for the 0-hr group. As the results will show, given that group differences in emotional effects were specific to consolidated Set 1, these would be hard to explain just on the grounds that for the 6-hr group the Day 8 combined test occurred late (instead of early) afternoon.

The experiment was controlled using DMDX (Forster & Forster, 2003). In Session 1, participants were exposed to 20 picture-novel word associations. On each trial, a target picture was presented on the screen together with one up-to five distractors, always from the same valence condition, while the corresponding novel word was both played in the headphones and printed in Courier New font 12 on the centre of the screen. In contrast to Richards and Blanchette (2004), the display was such that the printed word never obstructed any of the pictures (Fig. 1b). Participants had 20 seconds to decide which picture the novel word referred to and press the corresponding digit on the top row of the keyboard. As soon as the response was recorded, all distractors disappeared while the correct picture remained on the screen for another half a second, after which the next trial started. Each of the 20 associations were presented 21 times in blocks of 20 trials, with the number of distractors increasing

from one to three at Block 10 and from three to five at Block 17, and with the picture size getting smaller and smaller.[7] The correct picture was equiprobable at all locations and every distractor appeared the same number of times with every target. For the last block, the feedback was switched off to assess learning. In Session 2 (i.e., a week later) participants were first exposed to another set of 20 associations, following the same procedure as for the first set, but with different pictures and novel words. They were then tested on both sets of associations in the emotional Stroop, the pause detection and the picture-word association task, either immediately or after six hours of being awake (i.e., taking a nap was forbidden).

On each trial the emotional Stroop task required participants to identify as quickly as possible the colour in which the letter string presented was displayed. Four colours (blue, green, red, yellow) and thus four keys were used and all the 80 novel words were presented once in each colour. Participants had a maximum of 10 seconds to respond, followed by a 500-ms break. The 320 trials were spread over two 'floating' blocks within which trial order was randomized on each run. Before these, participants had 32 practice trials. In the pause detection task, participants had to decide as quickly as possible (by pressing one of two buttons) whether an artificial disruption, i.e., a 200-ms silence, was present at any location within each spoken stimulus. After 32 practice trials, the 80 novel words were presented, half of which contained a pause anywhere in the word. Participants had three seconds from stimulus onset to respond, followed by a one-second break. Item order was randomized on each run. Finally, in the picture-word association task, participants were tested on the 40 associations they had been exposed to, using five distractor pictures and no feedback, as on the final block of each exposure, and full randomization blind to valence and age of the item in memory.

4. Results

The data were examined using analyses of variance (ANOVAs) with participant as random factor. For picture-word association, fixed factors included group (0-hr, 6-hr), associations set/age

[7] The number of 20 exposures (plus one without feedback) was chosen taking into account the need to make the encoding phase salient enough that it would promote consolidation. Yet, it also had to be manageable, so that participants would not be too tired for the test phase—for half of them, the combined test followed immediately after exposure to Set 2. Twenty is half-way between the 16 exposures used by Clay et al. (2007) and the 24 exposures used by Geukes et al. (2015).

(first/seven-day old, second/same-day) and emotional valence (alarming, neutral). For pause detection and Stroop, a first ANOVA examined learning independently of emotional valence, by including only group and memory age (untrained, same-day, seven-day old). A second ANOVA included emotional valence on top of the other two factors. Chronometrical analyses were based on correct response latencies. From these, reaction times in the speeded tasks longer than 1,600 ms were rejected (1.57% in Stroop; 3.96% in pause detection). All accuracy/error analyses were carried out on arcsined proportions. As 'group' was a between-subject factor, only its interactive effects were considered. In pause detection, the analyses collapsed pause-present and pause-absent trials.

4.1 Immediate (initial acquisition) test

The 20 cycles of exposure were enough for accuracy in the **picture-word association task** to be near ceiling level on the test block that immediately followed each training session (Fig. 4a). Actually, as one can see on the right panels of Fig. 3, accuracy was near the ceiling very early on during the encoding phase (i.e., by Block 3), showing only a minimal drop (-1.1% on average) on Block 10 due to the increase from one to three distractor pictures. (A detailed analysis of the learning performance block-by-block is presented at Appendix B.) Consequently, the immediate test showed no effect of emotional valence (97.5% vs. neutral: 98.5%; $F(1,62) = 2.17$, $p > .14$), and apart from a marginal interaction due to Set 2 being 2.1% less accurate in the 6-hr (compared to the 0-hr) group ($F(1,62) = 2.84$, $p = .096$), no other effect or interaction approached significance (Fs < 1).

The disturbing nature of alarming trials was, nonetheless, clearly visible on latencies, which were on average 260 ms longer compared to neutral trials ($F(1,62) = 28.9$, $p < .00001$; Fig. 4b). In other words, our picture sets were working well. Except for a marginally significant interaction between group and emotional valence $F(1,62) = 3.34$, $p = .073$), due to a stronger valence effect in the 6-hr group (352 ms vs. 0-hr group: 173 ms), no other main or interactive effect approached significance (Fs$(1,62) \leq 1.10$, ps > .29).

4.2 Combined (consolidation) test

4.2.1 Picture-word association

At test, accuracy in the picture-word association task showed forgetting over the course of the week: associations learnt seven days before the test were not as well remembered as those learnt

minutes to hours before the test (82.6% vs. 96.2%; $F(1,62) = 65.33$, $p < .00001$; Fig. 4a). But with performance moving away from the ceiling over the course of the week, the impeding effect of alarming trials (e.g., 78.8% vs. neutral: 86.4%) was visible on the seven-day-old associations ($F(1,62) = 12.58$, $p < .0008$; same-day associations: $F < 1$). This pattern was supported by the presence of a significant interaction between memory age and valence ($F(1,62) = 10.64$, $p < .002$). The effect of valence on seven-day-old associations was numerically stronger when tested immediately after learning Set 2, instead of six hours later (10.9% vs. 4.4%, respectively). However, the two-way interaction of interest between group and valence was not significant ($F(1,62) = 1.24$, $p > .26$), nor was there any other effect or interaction (Fs < 1).

Latencies revealed a highly significant effect of valence ($F(1,62) = 25.86$, $p < .00001$; Fig. 4b), and although this effect was numerically stronger for same-day than for seven-day-old associations (505 ms vs. 397 ms), the interaction between memory age and valence was not significant ($F < 1$). Latencies confirmed the forgetting of seven-day-old associations relative to the same-day associations ($F(1,62) = 65.15$, $p < .00001$). This effect interacted with group ($F(1,62) = 14.30$, $p < .0004$) because, relative to training, the 6-hr group (i.e., tested six hours after learning Set 2) showed some forgetting also for same-day associations. A similar trend was visible – though not reliable – on the accuracy data. There was no other significant interaction (all Fs $(1,62) \leq 1.28$, ps > .26).

4.2.2 Pause detection

Latencies (Fig. 5a) showed no main effect of learning, or interaction with group ($Fs(2,124) \leq 1.61$, $p > .20$). Instead, they provided evidence that word emotional attributes need to consolidate to be operational: while words learnt on the day of the test showed a 4-ms nonsignificant advantage for the alarming condition ($F < 1$), the seven-day-old alarming words generated substantial and reliable interference (+30 ms) ($F(1,62) = 8.82$, $p < .005$). This change as a function of memory age was backed up by a significant two-way interaction with valence ($F(1,62) = 5.20$, $p < .03$). Also, there was no difference between the 0-hr group and the 6-hr group ($F < 1$), meaning that the interference observed after eight days had similar strength in both groups and was not yet present after six hours in the awake state following training. Error rates were low (3.7%) and showed no significant effect of

learning or valence, or interaction, even with group (all *F*s < 1, except for learning: $F(1,62) = 1.50$, $p > .22$).

To confirm that the emotional interference seen on seven-day-old items was triggered by the current trial, and was not a carry-over from the previous trial (see McKenna & Sharma, 2004), we checked that seven-day-old alarming *n* trials were not, for some unexpected reasons (trial order was fully randomized), preceded by more alarming than neutral *n-1* trials, and vice versa for neutral *n* trials. In this scenario, a slow attentional effect from trial *n-1* could account for the longer latencies on alarming *n* trials. But no, the proportions of combined trials were all similar to one another, and so a slow effect of this type cannot be the explanation (overall, alarming *n*: .48/.52 vs. neutral *n*: .50/.50; seven-day-old *n-1*, alarming *n*: .48/.52; neutral *n*: .49/.51; same-day *n-1*, alarming *n*: .48/.52; neutral *n*: .51/.49).

To produce a more stringent test of the fast nature of the emotional interference, we restricted our analysis to those *n* trials preceded by baseline or neutral *n-1* trials (for a similar approach, see Frings, Englebert, Wentura, & Bertmeitinger, 2010). These *n-1* trials, by definition, cannot contribute to a cumulative effect of negative emotion across trials that would surface on trial *n*. The results were unequivocal, with +28-ms interference on alarming (compared to neutral) seven-day-old items ($F(1,62) = 5.59$, $p < .03$), but no effect on same-day items (+2 ms, $F < 1$). Thus, there was nothing to suggest that the emotional interference observed on seven-day-old items originated from the previous trial.

To explore whether a slow effect from trial *n-1* could still be at work in the data (though evidently not as the driving force behind the interference), we looked at whether the age of the previous trial mattered for the emotional interference generated by seven-day-old trials. If, as the data indicated thus far, emotional attributes are operational only after consolidation, then, trial *n-1* would be most likely to contaminate trial *n*, if trial *n-1* also has had time to consolidate. The emotional interference vanished when trial *n-1* was also seven-day old (9 ms, $F < 1$), compared to when it was same-day or baseline (34 ms, $F(1,59) = 7.29$, $p = .009$), but the age of trial *n-1* x valence interaction failed to come out ($F < 1$). The consolidated representations evoked by trial *n-1* (irrespective of their emotional status) might have taken up resources needed to capture attention on trial *n*; alternatively, alarming *n-*

1 trials imposed their emotionality onto neutral *n* trials in such a way that the latter were as hard as alarming *n* trials.

4.2.3 Stroop

Latencies in the Stroop task (Fig. 5b) a priori showed no effect of learning or emotional valence, whatever the group (all *F*s(1,62) ≤ 1.06, *p*s > .30). Errors rates (4.9% overall), however, showed a hint of emotional interference from seven-day-old associations in the 0-hr group (5.0% vs. 3.4%; $F(1,31) = 2.90$, $p = .099$), but not in the 6-hr group (5.9% vs. 5.1%; $F < 1$), though the three-way interaction was not significant ($F(1,62) = 2.49$, $p > .11$; all other *F*s ≤ 1.63, *p*s > .20).

Given that pause detection showed signs that consolidated *n-1* trials influenced performance on *n* trials, we examined whether the absence of emotional Stroop, especially on seven-day-old associations, reflected a lack of sensitivity, or a contamination of trial *n* by trial *n-1*. Compared to pause detection, stimuli in the (visual) Stroop task came at a much faster pace (1,162 ms vs. 1,980 ms, on average). Hence, trial *n-1* had even more opportunity to influence performance on trial *n*. Besides, we know that the emotional Stroop sometimes evolves as a slow component affecting mostly the next trial.

4.2.3.1 Stroop and the age of trial *n-1* and trial *n*

We first looked at the role of memory age – irrespective of emotional valence – in the interplay between trial *n-1* and trial *n*. As illustrated in Fig. 6a, when the current and previous trials both tapped into seven-day-old associations, latencies on the current trial were significantly longer (+29 ms) compared to when trial *n-1* showed an item learnt earlier on that day or an unknown letter string (*F1*s (1,62) ≥ 12.18, *p*s < .0009). This pattern was confirmed by a significant interaction between the age of trial *n-1* and the age of trial *n* ($F(4,248) = 3.07$, $p < .02$), unaffected by group ($F < 1$), and by a main effect of trial *n-1* only when trial *n* was seven-day old ($F(2,124) = 8.90$, $p < .0003$; for new and same-day: *F*s < 1).

As shown in Fig. 6b, current seven-day-old trials *n* were matched in terms of latencies on trials *n-1* ($F < 1$). In other words, the interference observed on trial *n* originated from trial *n* (i.e., it was not a leftover from possible difference in performance on trials *n-1*). Two elements allow us to also reject

the possibility that this interference was a cumulative effect of emotion across trials: (1) for both seven-day old *n-1* trials, which gave rise to interference on seven-day old trials, and same-day *n-1* trials, which did not, the proportions of combined trials of each emotional crossing were all close to .25 (i.e., seven-day old: .23-.27; same-day: .23-.26); (2) the magnitude of the effect was unchanged whether trials *n-1* and *n* tapped into alarming associations, or neutral ones (34 ms vs. 43 ms; $F < 1$). Given the restriction of the effect to seven-day-old items, instead what we seem to have here is a new measure of memory consolidation, at the item-set level. Errors showed no trial-to-trial interaction ($F < 1$), whatever the group ($F(4,248) = 1.59$, $p > .17$).[8]

4.2.3.2 Neutralizing the influence of *n-1* to uncover emotions generated by *n*

Given the contaminating influence of trial *n-1* seen just by focusing on the age of these associations in memory, we revisited our emotional Stroop data with the following prediction in mind: if, as pause detection suggests, word emotional attributes take between six hours and seven days to be capable of capturing attention, then the most auspicious cases where an effect of emotional valence could be seen on trial *n* should be precisely those where trial *n-1* either activates an association that has not had time to consolidate (i.e., same-day items), or simply does not activate anything (i.e., baseline strings). And if the logic is correct, the interference should emerge only on seven-day-old (consolidated) items.

Fig. 7 presents the Stroop latencies as a function of the age of the association tapped into by the preceding trial *n-1*, separately for seven-day-old and same-day *n* trials. When trial *n-1* tapped into a seven-day-old association, trial *n* showed no sign of emotion effect, whether on the same-day or the seven-day-old items (both Fs < 1; interaction age of *n* by valence: $F < 1$). In contrast, when trial *n-1* tapped into a same-day association, and so could not emotionally contaminate trial *n*, the latter showed interference (+19 ms for alarming words) on seven-day-old items ($F(2,62) = 2.46$, $p = .06$, one-tailed), but nothing on same-day items ($F < 1$), exactly as predicted. This pattern was confirmed

[8] The same analysis run on pause detection data only showed what looks like a simple item-set effect (i.e., with no consolidation asymmetry): on average, latencies were 28 ms longer if trial *n-1* had been learnt as part of the same (as opposed to another) item-set, as trial *n*. This was supported by the interaction between the age of trial *n-1* and that of trial *n*, close to significance when baseline *n-1* and *n* trials were excluded ($F(1,60) = 3.66$, $p = .06$).

by the presence of a marginal interaction between the age of trial *n-1* and valence on seven-day-old *n* trials ($F(1,62) = 2.70$, $p = .10$). In other words, neutralizing the influence of *n-1* allowed us to somewhat uncover the same effect of consolidated emotions on trial *n*, as found in pause detection.

While the above logic predicted a similar result for trials preceded by unknown letter strings, these clearly had a special status, in that they appeared to impair lexical access on the following trial. This view is supported by the fact that latencies on trained items (i.e., same-day and seven-day old) were 11-ms shorter when preceded by an unknown baseline string as opposed to a learnt item (657 ms vs. 668 ms; $F(1,63) = 6.09$, $p < .02$; this is most visible on Fig. 7b). Under these circumstances, the emotional attributes of trial *n* would be activated too late or too weakly to have a measurable effect.

For the sake of completeness, we also examined whether, conversely, neutralizing the emotions generated by trial *n* allowed the consolidated alarming attributes of trial *n-1* to get through and monopolize attention (as in Frings et al., 2010). There was a trend for trial *n-1* to produce the strongest interference (+16 ms) when trial *n* was unconsolidated (as opposed to 0 ms for baseline and +6 ms for seven-day old), but this effect was not reliable ($F(1,63) = 1.44$, $p > .23$).

5. Discussion

Word emotional attributes take between six hours and seven days to be operational and capable of capturing attention. In two selective attention tasks (i.e., pause detection and emotional Stroop), novel word alarming (i.e., threat-related) attributes were found to impair participants' ability to categorize the word along the relevant dimension, not immediately or after six hours in the awake state following exposure, but only when tested after a week. In pause detection, the resulting interference was visible on latencies, whether the test was run immediately after learning the same-day associations or six hours later. In the emotional Stroop hints of interference from consolidated emotions were seen in the errors, at least when exposure to the same-day associations occurred immediately before the test. This interference was seen also on Stroop latencies, when the association targeted by the previous trial had not had time to consolidate and thus could not contaminate the current trial (i.e., a seven-day-old trial *n* preceded by a same-day trial *n-1*).

Though the Stroop results are more complex due to the influence of the previous trial, both tasks show that novel words need to consolidate in order to acquire their emotional status. That the same-

day items showed emotional interference in neither task makes it unlikely that such a null effect would reflect a lack of sensitivity**,** later overcome thanks to memory consolidation making the underlying episodic details more salient. If this was the explanation, picture-word association which, in effect, contains an episodic reminder of the exposure, would have shown a larger threat reaction to seven-day-old than to same-day associations (e.g., Payne et al., 2006). This is not what we found: latencies (unconstrained by the ceiling) showed a reaction to threat of similar magnitude for same-day and seven-day-old associations. Thus, taken together, these results indicate that the link between the new word forms and their emotional semantics were just not yet available minutes-to-hours after encoding.

As such, these findings cast doubt on the idea that the Stroop interference found immediately and with far fewer exposures by both Richard and Blanchette (2004; see also Blanchette & Richards, 2013) and Sharma and Money (2010) would reflect a language-mediated reaction to threat. As we argued in the Introduction, the training procedure used these studies had little control over how much of the word versus the picture was encoded on each trial, and this could easily have varied depending on the nastiness of the image and the anxiety of the participants. Given the immediacy of their effect, it is more likely to reflect modulation of attention by emotions at the time of encoding, rather than threat reactivation during emotional Stroop (i.e., at retrieval). In contrast, when the training procedure ensures that participants pay attention to both ends of each association, as is the case in the present study, what we find is that the acquisition of word emotional attributes requires offline consolidation. In that respect, our results corroborate the step-like function in the acquisition of lexical semantics found by Clay et al., (2007) and Geukes et al. (2015) using the picture-word interference and the colour Stroop paradigm. They also add to a decade of studies showing all kinds of offline, mostly sleep-related, changes in the representation of newlylearnt words (e.g., Bakker, Takashima, van Hell, Janzen, & McQueen, 2015; Bowers et al., 2005; Dumay, 2016; Dumay & Gaskell, 2007, 2012; Gaskell & Dumay, 2003; Leach & Samuel, 2007; Tamminen, Davis, & Rastle, 2015; see Kapnoula & McMurray, 2016, for an opposing view).

In the present study, the 'hijacking' effect of threat on attention showed both its fast and slow facets. However, for the latter, this was more in transparency, so to speak. In pause detection the

interference from consolidated emotion was strongest when the association tapped by trial *n-1* had not had time to consolidate (i.e., same day) or just did not exist (i.e., baseline). But, conversely, restricting the analysis to those *n* trials preceded by a non-emotional (i.e., baseline or neutral) trial *n-1* showed the exact same pattern as found overall, with interference only for *n* trials that tapped into a seven-day (i.e., consolidated) association. A similar dynamic was found in the Stroop task, though here interference from consolidated emotion was visible only after neutralizing the influence of trial *n-1*. That the age of trial *n-1* dictates the emergence of emotional interference on trial *n* shows a carryover from trial *n-1*. But the fact that the interference on trial *n* is unshaken when trial *n-1* is nonemotional tells us that the effect is triggered by the current trial *n*. In other words, a true capture of attention by the current word *can* be obtained. These data also corroborate, here using a within-item manipulation, the results by Frings et al. (2010) suggesting the co-occurrence of a fast and a slow effect, and not just one or the other, on consecutive trials.

Another element of the emotional Stroop effect in the present study is its disappearance when trial *n-1* showed a nonword that was not part of the exposure (i.e., a baseline item). In this case, colour classification was faster than anywhere else. This suggests that participants either did not engage as much in reading the subsequent string or that response to trial *n-1* somehow sped-up latencies to trial *n*, with the result that the colour response was produced before attention was captured. This aspect of the data is reminiscent of list effects in word recognition, whereby the influence of lexical variables vanishes when words are presented intermixed with pseudowords (see Traficante & Burani, 2014, for a review). Only, what is a word or a pseudoword in the present study is determined just by whether or not the string was part of the exposure. Typically, list-composition effects have been explained in one of two ways: the failure to achieve lexical access on the preceding trial shifts the processing mode to one that puts less emphasis on lexical access and more on transcoding from print-to-sound based on sublexical mappings (e.g., Baluch & Besner, 1991; Monsell, Patterson, Graham, Hughes, & Milroy, 1992; Peereman & Content, 1995); or participants simply tend to homogenize their response times, taking into account their speed on the preceding trial (e.g., Kinoshita, Mozer, & Forster, 2011; Lupker, Brown, & Colombo, 1997; Rastle, Kinoshita, Lupker, & Coltheart, 2003). The fact that latencies to trial *n* overall here were not any faster than for the other conditions would make it hard to

explain the disappearance of the emotional interference in terms of participants adapting their speed. But whatever the underlying mechanism, suppression of emotional Stroop by an unknown word on the preceding trial is another item for the list of arguments against the automatic nature of the phenomenon (see Bertels & Kolinsky, 2015, for a recent discussion).

While both selective attention tasks show that memory consolidation plays a key role in the acquisition of word emotional attributes, alarming associations returned a poorer performance than neutral associations at both the immediate and combined explicit tests. At first glance, this may appear to go against the well-documented memory advantage for emotionally-loaded information: faces, objects, scenes and words are usually better remembered if they are arousing and/or negatively valenced, not just because they may attract attention, but also because they engage affect-specific brain circuitries that modulate activity of the medial temporal lobe (for reviews, see Buchanan & Adolphs, 2002; Kensinger, 2007; Hamann, 2001). However, this would be losing sight of the fact that our test trials were not equated for how much emotional reaction they would trigger on the fly: because we did not want the immediate test to dampen our manipulation, alarming associations were always tested against alarming picture distractors, and vice versa for neutral associations. Thus, the poorer performance with alarming associations most likely reflects a reduced ability to perform the task simply because of emotional disturbance (see Kensinger & Corkin, 2003, for similar effect on working memory). Neither the learning rate during exposure (Fig. 3 and also Appendix B), nor the rate of forgetting captured by the difference between same-day and seven-day-old associations at the combined test (Fig. 4) showed a difference in the long-term fate of alarming and neutral associations.

In that respect, our results are in broad agreement with the few studies that looked for a role for emotion in the formation of associations between 'free-standing' stimuli. In these, the effect of emotion on memory appeared to be a more subtle one, only modulating the influence of other factors. Murray and Kensinger (2012) asked participants to encode pairs of unrelated words (e.g., card-mouse) either by visualizing them as two separate entities, or as single integrated associations. They found a dissociation such that long-term memory for neutral-emotional pairs (tested after 30 minutes) did not benefit as much as neutral-only pairs from encoding in an integrative, as opposed to nonintegrative, fashion. A similar memory head-start was found in a recent nap study by Alger and Payne (2016).

They were looking at whether the benefit of sleep on relational (i.e., mediated) memories of faces involved in pairs that shared a common object was itself modulated by the object's emotional attributes. The authors found no effect of emotion on relational memory. In contrast, the face-object pairs originally presented at encoding showed a dissociation: in a similar way as with the type of encoding in Murray and Kensinger, the neutral pairs benefitted more from the stabilizing effect of sleep.

On the front of emotions as well, the only evidence that we found for the idea that exposure to alarming pictures had sensitized participants lied in the Stroop task. This task showed more errors on alarming than neutral seven-day-old associations, but only when these were tested immediately after exposure to Set 2, instead of after six hours. As it was only marginally significant, this effect awaits replication. Nonetheless, it already suggests that language-mediated effects of emotion on attention may be best captured after subjecting participants to some emotional disturbance (for instance, by showing them some of the IAPS pictures). An implication of this is that if a retest no longer shows an effect of emotion that was found at an earlier point, one should ask whether forgetting has occurred, or, alternatively, whether some aspect of the protocol sensitized participants before the immediate test, but not before the retest.

Besides the need for word emotional attributes to consolidate before they can hijack attention, the other main finding of this study is the interference observed in the Stroop task, irrespective of emotion, when the current and preceding trials activated consolidated words. This is as if memory consolidation gave novel words learnt within the same set the ability to recognize one another across trials. One possible explanation is that consolidation (mostly sleep-induced) over the course of the week boosts access to these individual memories by strengthening either their episodic representations in the hippocampus, or their neocortical counterparts (via neural replay), or both. Reactivation of two of these consolidated traces by consecutive trials would be sufficient to momentarily disturb the frail balance between the current task (i.e., classifying the colour) and the task performed at encoding while learning the items (i.e., word reading), in favour of the latter. In other words, the accumulation of episodic reminders over consecutive trials would make it hard *not* to fall back on the task set that was consolidated together with the items. The proactive control/task conflict model recently proposed

by Kalanthroff, Avnit, Henik, Davelaar and Usher (2015; Kalanthroff, Henik, Derakshan, & Usher, 2016), if sensitive to variations of memory strength, may provide the right framework.

Kalanthroff et al. emphasize that the main problem in selective attention is maintaining task demands. In their model, these are controlled both top-down (e.g., by the instructions) and bottom-up (e.g., by the stimulus properties). When, during colour naming, word reading is inadvertently activated, because, say, the stimulus is orthotactically legal (as opposed to a series of unpronounceable strings), task conflict occurs, with the irrelevant task suppressing the relevant task response. Besides evidence from the colour-word Stroop task showing that this can happen even when the colour response is semantically congruent, support for this account can be found in a recent priming study by Sharma (in press). In this study, words studied beforehand were found to produce slower colour naming latencies only if trial *n-1* was also part of the study list. Sharma interprets this reverse Gratton sequential modulation (Gratton, Coles, & Donchin, 1992) in terms of increased task conflict. In the Kalanthroff et al. model, this is the only way to generate interference via a trial-to-trial cumulative process. As the interference found for consolidated items in the present study shows a similar dependence on trial *n-1*, task conflict induced by an accumulation of episodic reminders could be behind the effect here as well.

Another possibility is that systems consolidation (via neural replay) strengthens both hippocampal and neocortical associations between the items of the same set. When two items of the same consolidated set are presented in direct succession, their strengthened association could blur the temporal distinctiveness between the two events. This would lead participants to at first cancel their response to trial *n*, because they have already responded on trial *n-1*. A variant could also be that with this special bond between the two trials, trial *n* may reactivate the response given on trial *n-1*, leading to a *response* conflict in 75 percent of the trials – there were four colours in the Stroop task. This account makes predictions testable within the current dataset: whereas consecutive seven-day-old trials should produce interference on trial *n* when the colour is not repeated across trials (i.e., unrepeated response), they should produce facilitation instead when the colour *is* repeated (i.e., repeated response).

Obviously, splitting the data between colour-unrepeated and colour-repeated trials showed a massive response repetition priming effect in all conditions, with latencies dropping from 715 ms to 508 ms, on average, between unrepeated and repeated colour trials. With latencies so close to the floor on repeated trials, it is hard to see whether consolidated *n-1* and *n* trials produced the expected facilitation in this case (-6 ms in the right direction, $t(63) = -.02$). To go around this, we computed the by-participant correlation between the magnitude of response repetition priming and the difference between seven-day-old *n* trials preceded by seven-day-old vs. same-day/baseline *n-1* trials. The obtained correlation ($r(64) = .07$, *ns*) did not help adjudicating between task conflict, which predicts a negative correlation due to colour repetition squashing down the interference seen in the unrepeated case, and response conflict, which assumes that in this case what should be squashed is a facilitation effect. A median-split gave us an answer: participants who showed the weakest response repetition priming from trial *n-1* to trial *n* (in all other conditions) exhibited a 30-ms facilitation when both trials were seven-day old, as compared to when trial *n-1* was not ($t(31) = -1.93$, $p = .063$; and adding two more participants on this side of the split confirmed the tendency: $t(33) = -2.13$, $p < .05$; Fig. 6a). Such a flip, from inhibition when the colour is unrepeated to facilitation when it is repeated across trials, is evidence for a response conflict. In other words, the observed cross-trial interaction most likely originates from the offline strengthening the indirect associations between the words of the set. Interestingly, in contrast to all previous studies, which used overlapping associations of the type "A-B and B-C" (e.g., Alger & Payne, 2016; Lau, Tucker, & Fishbein, 2010), here the binding element resides only in the spatio-temporal context of the session, in particular the fact that pictures with the same valence were all used as distractors for one another.[9]

Thus, to sum up, a true emotional Stroop interference (i.e., uncontaminated by stimulus idiosyncrasies) can be obtained from newly-learnt made-up words, as long as the test occurs after a consolidation interval, which most likely needs to include sleep. After consolidation, word emotional attributes are able to both hijack attention on the current trial and modulate what happens on the

[9] We dedicate this finding to the memory of Paul Bertelson (1926-2008).

subsequent trial, but the words themselves, as long as they were learnt as part of the same set, also appear to recognize one another.

Author contributions

ND designed and prepared the experiment, analysed the data and wrote the paper. DS contributed to the design and stimulus preparation, and bounced ideas throughout. NK and SA helped with stimulus selection and tested all participants.

Acknowledgements

We thank Jacqueline Aldridge for lending us her voice, Arty Samuel and two anonymous reviewers for commenting on an earlier draft, and Valerie Streak for proofreading.

References

Alger, S.E., & Payne, J.D. (2016). The differential effects of emotional salience on direct associative and relational memory during a nap. *Cognitive, Affective, & Behavioral Neuroscience*, 16, 1150-1163.

Bakker, I., Takashima, A., van Hell, J.G., Janzen, G., & McQueen, J.M. (2015). Tracking lexical consolidation with ERPs: Lexical and semantic-priming effects on N400 and LPC responses to newly-learned words. *Neuropsychologia*, 79, 33-41.

Balota, D.A., Cortese, M.J., Hutchison, K.A., Neely, J.H., Nelson, D., Simpson, G.B., & Treiman, R. (2002). *The English Lexicon Project: A Web-based repository of descriptive and behavioral measures for 40,481 English words and non-words.* Available at http://elexicon.wustl.edu

Baluch, B., & Besner, D. (1991). Strategic use of lexical and nonlexical routines in visual word recognition: Evidence from oral reading in Persian. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17, 252-259.

Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M.J., & van Ijzendoorn, M.H. (2007). Threat-related attentional bias in anxious and non-anxious individuals: A meta-analytic study. *Psychological Bulletin*, 133, 1-24.

Bertels, J., & Kolinsky, R. (2015). Disentangling fast and slow attentional influences of negative and taboo spoken words in the emotional Stroop paradigm. *Cognition and Emotion*, 30, 1137-1148.

Bisby, J.A., & Burgess, N. (2014). Negative affect impairs associative memory but not item memory. *Learning & Memory*, 21, 760-766.

Bishop, S.J., Duncan, J., & Lawrence, A.D. (2004). State anxiety modulation of the amygdala response to unattended threat-related stimuli. *Journal of Neuroscience*, 24, 10364-10368.

Bishop, S.J., Duncan, J., Brett, M., & Lawrence, A.D. (2004). Prefrontal cortical function and anxiety: controlling attention to threat-related stimuli. *Nature Neuroscience*, 7, 184-188.

Blanchette, I., & Richards, A. (2013). Is Emotional Stroop Interference Linked to Affective Responses? Evidence From Skin Conductance and Facial Electromyography. *Emotion*, 13, 129-138.

Bowers, J.S., Davis, C.J., & Hanley, D.A. (2005). Interfering neighbours: The impact of novel word learning on the identification of visually similar words. *Cognition*, 97, B45-B54.

Buchanan, T.W., & Adolphs, R. (2002). The role of the human amygdala in emotional modulation of long-term declarative memory. In S. Moore & M. Oaksford (Eds.), *Emotional Cognition: From Brain to Behavior* (pp. 9-34). Amsterdam: John Benjamins Publishing.

Burt, J.S. (2002). Why do non-colour words interfere with colour naming? *Journal of Experimental Psychology: Human Perception and Performance*, 28, 1019-1038.

Camblats, A.M., & Mathey, S. (2016). The effect of orthographic and emotional neighbourhood in a colour categorization task. *Cognitive Processing*, 1, 115-122.

Caparos, S., & Blanchette, I. (2014). Emotional Stroop interference in trauma-exposed individuals: A contrast between two accounts. *Consciousness and Cognition*, 28, 104-112.

Carretié, L. (2014). Exogenous (automatic) attention to emotional stimuli: A review. *Cognitive, Affective & Behavioral Neuroscience*, 14, 1228-1258.

Cisler, J.M., & Koster, E.H. (2010). Mechanisms of attentional biases towards threat in anxiety disorders: An integrative review. *Clinical Psychology Review*, 30, 203-216.

Cisler, J.M., Wolitzky-Taylor, K.B., Adams Jr, T.G., Babson, K.A., Badour, C.L., & Willems, J.L. (2011). The emotional Stroop task and posttraumatic stress disorder: A meta-analysis. *Clinical Psychology Review*, 31, 817-828.

Clay, F., Bowers, J.S., Davis, C.J., & Hanley, D.A. (2007). Teaching adults new words: The role of practice and consolidation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 33, 970-976.

Cox, W.M., Fadardi, J.S., & Pothos, E.M. (2006). The addiction-Stroop test: Theoretical considerations and procedural recommendations. *Psychological Bulletin*, 132, 443-476.

Cunningham, T.J., Crowell, C.R., Alger, S.E., Kensinger, E.A., Villano, M.A., Mattingly, S.M., & Payne, J.D. (2014). Psychophysiological arousal at encoding leads to reduced reactivity but enhanced emotional memory following sleep. *Neurobiology of Learning and Memory*, 114, 155-164.

Cutler, A. (1981). Making up materials is a confounded nuisance, or: Will we be able to run any psycholinguistic experiments at all in 1990? *Cognition*, 10, 65-70.

Dalgleish, T. (2005). Putting some feeling into it—the conceptual and empirical relationships between the classic and emotional Stroop tasks: Comment on Algom, Chajut, and Lev (2004). *Journal of Experimental Psychology: General*, 134, 585-591.

Davis, C.J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods*, 37, 65-70.

Dumay, N. (2016). Sleep not just protects memories against forgetting, it also makes them more accessible. *Cortex*, 74, 289-296.

Dumay, N., & Gaskell, M.G. (2007). Sleep-associated changes in the mental representation of spoken words. *Psychological Science*, 18, 35-39.

Dumay, N., & Gaskell, M.G. (2012). Overnight lexical consolidation revealed by speech segmentation. *Cognition*, 123, 119-132.

Easterbrook, J.A. (1959). The effect of emotion on cue-utilization and the organization of behavior. *Psychological Review*, 66, 183-201.

Eysenck, M.W., Derakshan, N., Santos, R., & Calvo, M.G. (2007). Anxiety and cognitive performance: attentional control theory. *Emotion*, 7, 336-353.

Epp, A.M., Dobson, K.S., Dozois, D. J., & Frewen, P.A. (2012). A systematic meta-analysis of the Stroop task in depression. *Clinical Psychology Review*, 32, 316-328.

Forster, K.I., & Forster, J.C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods*, 35, 116-124.

Frings, C., Englert, J., Wentura, D., & Bermeitinger, C. (2010). Decomposing the emotional Stroop effect. *Quarterly Journal of Experimental Psychology*, 63, 42-49.

Gaskell, M.G., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition*, 89, 105-132.

Geukes, S., Gaskell, M.G., & Zwitserlood, P. (2015). Stroop effects from newly learned colour words: effects of memory consolidation and episodic context. *Frontiers in Psychology*, 6:278.

Goldstein, A.N., & Walker, M.P. (2014). The role of sleep in emotional brain function. *Annual Review of Clinical Psychology*, 10, 679-708.

Gratton, G., Coles, M.G., & Donchin, E. (1992). Optimizing the use of information: Strategic control of activation of responses. *Journal of Experimental Psychology: General*, 121, 450-480.

Hamann, S. (2001). Cognitive and neural mechanisms of emotional memory. *Trends in Cognitive Sciences*, 5, 394-400.

Hu, P., Stylos-Allan, M., & Walker, M.P. (2006). Sleep facilitates consolidation of emotional declarative memory. *Psychological Science*, 17, 891-898.

Joormann, J. (2010). Cognitive inhibition and emotion regulation in depression. *Current Directions in Psychological Science*, 19, 161-166.

Kalanthroff, E., Avnit, A., Henik, A., Davelaar, E.J., & Usher, M. (2015). Stroop proactive control and task conflict are modulated by concurrent working memory load. *Psychonomic Bulletin & Review*, 22, 869-875.

Kalanthroff, E., Henik, A., Derakshan, N., & Usher, M. (2016). Anxiety, emotional distraction, and attentional control in the Stroop task. *Emotion, 16,* 293-300.

Kapnoula, E.C., & McMurray, B. (2016). Newly learned word-forms are abstract and integrated immediately after acquisition. *Psychonomic Bulletin and Review*, 23, 491-499.

Kensinger, E.A., & Corkin, S. (2003). Effect of negative emotional content on working memory and long-term memory. *Emotion*, 3, 378-393.

Kensinger, E.A. (2007). Negative emotion enhances memory accuracy: Behavioral and neuroimaging evidence. *Current Directions in Psychological Science*, 16, 213-218.

Kensinger, E.A. (2009). Remembering the details: Effects of emotion. *Emotion Review*, 1, 99-113.

Khanna, M.M., Badura-Brack, A.S., McDermott, T.J., Shepherd, A., Heinrichs-Graham, E., Pine, D.S., Bar-Haim, Y., & Wilson, T.W. (2015). Attention training normalises combat-related post-traumatic stress disorder effects on emotional Stroop performance using lexically matched word lists. *Cognition and Emotion*, 30, 1521-1528.

Kinoshita, S., Mozer, M.C., & Forster, K.I. (2011). Dynamic adaptation to history of trial difficulty explains the effect of congruency proportion on masked priming. *Journal of Experimental Psychology: General*, 140, 622-636.

Kinoshita, S., De Wit, B., & Norris, D. (2017). The magic of words reconsidered: Investigating the automaticity of reading color-neutral words in the Stroop task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43, 369-384.

Klein, G.S. (1964). Semantic power measured through the interference of words with color-naming. *The American Journal of Psychology*, 77, 576-588.

Lang, P.J., Bradley, M.M., & Cuthbert, B.N. (2008). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. *Technical Report A-8.* University of Florida, Gainesville, FL.

Larsen, R.J., Mercer, K.A., & Balota, D.A. (2006). Lexical characteristics of words used in emotional Stroop experiments. *Emotion*, 6, 62-72.

Lau, H., Tucker, M., & Fishbein, W. (2010). Daytime napping: Effects on human direct associative and relational memory. *Neurobiology of Learning and Memory*, 93, 554-560.

Leach, L., & Samuel, A.G. (2007). Lexical configuration and lexical engagement: When adults learn new words. *Cognitive Psychology*, 55, 306-353.

Lupker, S.J., Brown, P., & Colombo, L. (1997). Strategic control in a naming task: Changing routes or changing deadlines? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 570-590.

MacLeod, C.M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, 109, 163-203.

Magnuson, J.S., Tanenhaus, M.K., Aslin, R., & Dahan, D. (2003). The microstructure of spoken word recognition: Studies with artificial lexicons. *Journal of Experimental Psychology: General*, 132, 202-227.

Mattys, S.L., & Clark, J.H. (2002). Lexical activity in speech processing: Evidence from pause detection. *Journal of Memory and Language*, 47, 343-359.

McKenna, F.P., & Sharma, D. (2004). Reversing the emotional Stroop effect reveals that it is not what it seems: The role of fast and slow components. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 382-392.

Monsell, S., Patterson, K.E., Graham, A., Hughes, C.H., & Milroy, R. (1992). Lexical and sublexical translation of spelling to sound: Strategic anticipation of lexical status. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 452-467.

Monsell, S., Taylor, T.J., & Murphy, K. (2001). Naming the color of a word: Is it responses or task-sets that compete? *Memory and Cognition*, 29, 137-151.

Murray, B.D., & Kensinger, E.A. (2012). The effects of emotion and encoding strategy on associative memory. *Memory & Cognition*, 40, 1056-1069.

Nishida, M., Pearsall, J., Buckner, R.L., & Walker, M.P. (2009). REM sleep, prefrontal theta, and the consolidation of human emotional memory. *Cerebral Cortex*, 19, 1158-1166.

Öhman, A. (2005). The role of the amygdala in human fear: Automatic detection of threat. *Psychoneuroendocrinology*, 30, 953-958.

Payne, J.D., & Kensinger, E.A. (2010). Sleep's role in the consolidation of emotional episodic memories. *Current Directions in Psychological Science*, 19, 290-295.

Payne, J.D., Chambers, A.M., & Kensinger, E.A. (2012). Sleep promotes lasting changes in selective memory for emotional scenes. *Frontiers in Integrative Neuroscience*, 6:108.

Payne, J.D., Kensinger, E.A., Wamsley, E.J., Spreng, R.N., Alger, S.E., Gibler, K., Schacter, D.L., & Stickgold, R. (2015). Napping and the selective consolidation of negative aspects of scenes. *Emotion*, 15, 176-186.

Payne, J.D., Swanberg, K., Stickgold, R., & Kensinger, E.A. (2008). Sleep preferentially enhances memory for emotional components of scenes. *Psychological Science*, 19, 781-788.

Peereman, R., & Content, A. (1995). Neighborhood size effect in naming: Lexical activation or sublexical correspondences? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 409-421.

Pergamin-Hight, L., Naim, R., Bakermans-Kranenburg, M.J., van Ijzendoorn, M.H., Bar-Haim, Y. (2015). Content specificity of attention bias to threat in anxiety disorders: A meta-analysis. *Clinical Psychology Review*, 35, 10-18.

Phaf, R.H., & Kan, K.J. (2007). The automaticity of emotional Stroop: A meta-analysis. *Journal of Behavior Therapy and Experimental Psychiatry*, 38, 184-199.

Pratto, F., & John, O.P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology*, 61, 380-391.

Rastle, K., Kinoshita, S., Lupker, S.J., & Coltheart, M. (2003). Cross-task strategic effects. *Memory and Cognition*, 31, 867-876.

Richards, A., & Blanchette, I. (2004). Independent manipulation of emotion in an emotional Stroop task using classical conditioning. *Emotion*, 4, 275-281.

Sharma, D. (in press). Priming can affect naming colours using the study-test procedure. Revealing the role of task conflict. *Acta Psychologica.*

Sharma, D., & Money, S. (2010). Carryover effects to addiction-associated stimuli in a group of marijuana and cocaine users. *Journal of Psychopharmacology*, 24, 1309-1316.

Stroop, J.R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18, 643-662 (Reprinted in 1992 in the *Journal of Experimental Psychology: General*, 121, 15-23).

Takashima A., Nieuwenhuis, I.L., Jensen, O., Talamini, L.M., Rijpkema, M., & Fernández, G. (2009). Shift from hippocampal to neocortical centered retrieval network with consolidation. *Journal of Neuroscience*, 29, 10087-10093.

Tamminen, J., Davis, M. H., & Rastle, K. (2015). From specific examples to general knowledge in language learning. *Cognitive Psychology*, 79, 1-39.

Tamminen, J., Lambon Ralph, M., & Lewis, P.A. (2013). The role of sleep spindles and slow-wave activity in integrating new information in semantic memory. *Journal of Neuroscience*, 33, 15376-15381.

Traficante, D., & Burani, C. (2014). List context effects in languages with opaque and transparent orthographies: A challenge for models of reading. *Frontiers in Psychology*, 5:1023.

van Casteren, M., & Davis, M.H. (2007). Mix, a program for pseudorandomization. *Behavior Research Methods*, 38, 584-589.

Wagner, U., Gais, S., & Born, J. (2001). Emotional memory formation is enhanced across sleep intervals with high amounts of rapid eye movement sleep. *Learning and Memory*, 8, 112-119.

Williams, J.M.G., Mathews, A., & MacLeod, C. (1996). The Emotional Stroop task and psychopathology. *Psychological Bulletin*, 120, 3-24.

Williams, J.M.G., Watts, F.N., MacLeod, C., & Mathews, A. (1997). *Cognitive Psychology and Emotional Disorders (2nd ed.).* Chichester, UK: Wiley.

Wixted, J.T., & Cai, D.J. (2014). Memory consolidation. In K. Ochsner & S. Kosslyn (Eds.), *Oxford Handbook of Cognitive Neuroscience* (pp. 1-59). New York: Oxford University Press.

Yiend, J. (2010). The effects of emotion on attention: A review of attentional processing of emotional information. *Cognition and Emotion*, 24, 3-47.

Appendix A: Experimental materials.

IAPS pictures by ID number (plus valence and arousal in the population tested): 2900 (2.75; 6.43), 3030 (1.68; 7.73), 3220 (2.33; 6.6), 3350 (2.03; 7.30), 9910 (2.15; 7.25), 6212 (1.75; 7.73), 9040 (1.40; 7.80), 6510 (2.45; 7.75), 6570 (1.88; 7.48), 6821 (2.28; 7.35), 2840 (5.18; 4.43), 2200 (5.15; 4.85), 4571 (6.20; 4.70), 2351 (6.40; 4.68), 7595 (5.05; 4.33), 2410 (4.73; 5.43), 2620 (6.63; 3.38), 7285 (5.8; 4.00), 7550 (5.10; 4.23), 2870 (6.38; 4.35), 7380 (2.40; 7.20), 9006 (2.53; 6.88), 6260 (2.68; 7.93), 9050 (2.05; 7.35), 9140 (1.68; 6.93), 9253 (1.40; 8.03), 9410 (1.10; 8.15), 9560 (3.70; 5.40), 9433 (1.28 7.93), 9520 (2.35; 6.58), 2702 (5.08; 5.05), 5920 (5.15; 5.63), 7050 (4.85; 4.73), 7620 (5.85; 4.13), 1670 (6.15; 3.80), 2381 (6.08; 3.60), 9070 (4.53; 4.38), 1560 (5.30; 6.10), 2850 (5.68; 3.85), 9700 (4.65; 5.33). Note: The first 20 ID numbers refer to one set, whereas the remaining ID numbers refer to the other set. In each set, the first ten ID numbers refer to the alarming pictures.

Novel words: beegue, biente, blece, bligue, brabb, cewth, cieff, clauv, cluic, dreubb, drott, dweubb, dweugg, dwooph, frarsh, gawsh, gealte, ghlell, ghrudd, ghwaub, ghweep, ghwieg, ghwive, ghworc, ghylge, glaumb, gnaint, gnolck, gnoolt, gnurgn, greul, grezz, guegn, guilmn, gwaumn, hirnt, jarlt, kandge, knirck, knylgn, krarce, kreib, krewt, kridd, kriet, kurpe, lirlt, neilte, nunck, phalmb, phrirn, phrurt, psamph, pseum, psooce, quoarv, raibb, reegue, rhalp, ruigue, ruithe, shemph, shraum, shrurg, skorck, smawse, smeuve, smieve, smoone, speup, swaish, thilmb, thwass, thweap, thwiec, tultch, werdge, wheegg, wuilte, yaike.

Appendix B: Analysis of performance during encoding.

The increase in the number of distractor pictures from 1 to 3 on Block 10, and from 3 to 5 on Block 17 during the two encoding phases had the expected effect of shaking performance (see Fig. 3). On Block 10 latencies increased by 1,047 ms and accuracy dropped by 1.1%, and both of these effects were significant ($F(1,62) = 468.70$, $p < .0001$; $F(1,62) = 5.25$, $p < .03$, respectively). On Block 17, accuracy did not change significantly relative to Block 16 (+0.7%; ($F(1,62) = 1.16$, $p > .28$), but latencies again showed a substantial lengthening, of 448 ms ($F(1,62) = 113.72$, $p < .0001$). Given these expected changes in performance, accuracy and latencies during training were analysed separately for each of the three resulting time-windows. As for the test data, the ANOVAs that were carried out included group (0-hr, 6-hr), set (Set 1, Set 2) and emotional valence (alarming, neutral), but also now block (1-9, 10-16, and 17-20, respectively).

Blocks 1 to 9. The first time-window showed that accuracy improved from one block to the next until it reached a plateau near ceiling level ($F(8,488) = 126.5$, $p < .0001$); that it was poorer for alarming than neutral associations ($F(1,61) = 11.7$, $p < .002$); and that over the first two blocks the 6-hr group was less accurate with Set 2 than with Set 1, as reflected in the three-way interaction ($F(8,488) = 5.17$, $p < .0001$). Overall, accuracy for the 6-hr group did not differ from that of the 0-hr group ($F(1,61) = 1.95$, $p > .16$), and the valence effect was not significantly modulated by any of the other factors, whether individually or synergistically ($Fs \leq 1.51$, $ps > .15$). Latencies showed similar signs of improvement across blocks ($F(8,488) = 104.1$, $p < .0001$) and a practice effect from Set 1 to Set 2 ($F(1,61) = 28.9$, $p < .0001$) that shrank as performance improved (set x block: $F(8,488) = 20.7$, $p < .0001$). They also confirmed the poorer performance on alarming compared to neutral associations ($F(1,61) = 28.8$, $p < .0001$) and the absence of any modulation of this effect by any of the other factors ($Fs \leq 1.37$, $ps > .24$). Apart from a two-way interaction between block and group ($F(8,488) = 2.85$, $p < .005$) due to the 6-hr group being more conservative at the start of each session, no other effect or interaction approached significance ($Fs \leq 2.02$, $ps > .16$).

Blocks 10 to 16. During the second time-window, the 6-hr group performed better than the 0-hr group, as shown by a main effect of group on accuracy ($F(1,62) = 4.65$, $p < .04$) in the absence of a

trade-off with latencies ($F < 1$). Participants were also more conservative during their first encoding session, as indicated by their more accurate, but slower performance on Set 1 than on Set 2 (accuracy: $F(1,62) = 10.13$, $p < .003$; latencies: $F(1,62) = 6.30$, $p < .02$). The effect of valence here was significant only on latencies ($F(1,62) = 34.76$, $p < .0001$). With performance so close to the ceiling, accuracy only showed a trend in the same direction during exposure to Set 2 for the 0-hr group, which was picked up by the three-way interaction ($F(1,62) = 3.90$, $p = .053$). Apart from an effect of block on latencies, with participants recovering from the increase in number of distractors on Block 10 ($F(6,372) = 11.97$, $p < .0001$), no other effect or interaction approached significance (all $Fs \leq 1.21$, $ps > .30$).

Blocks 17 to 20. During the last time-window, Set 1 still showed a more accurate, but slower performance than for Set 2, though here only the difference in accuracy was significant ($F(1,62) = 7.39$, $p < .009$). The 6-hr group was only marginally more accurate than the 0-hr group ($F(1,62) = 3.55$, $p = .064$), and still without any trade-off with latencies ($F < 1$). The valence effect was significant on latencies ($F(1,61) = 15.57$, $p < .0003$), but also marginally significant on accuracy ($F(1,62) = 3.11$, $p = .082$), which contrasts with the preceding window. As indicated by the presence of a significant three-way interaction between block, set, and valence ($F(3,186) = 2.92$, $p < .04$), this was due to a drop in performance on alarming associations in the final block of exposure to Set 2. Apart from a three-way interaction close to significance ($F(3,183) = 2.54$, $p = .058$), due to the disappearance of the valence effect on the final block of both sessions in the 6-hr group, no other effect or interaction approached significance on latencies ($Fs(3,183) \leq 1.67$, $p > .17$).

Figure captions

Fig. 1. a) A negative free-viewing conditioning trial in the Richards and Blanchette (2004) study. The word appeared only after 500 ms into the picture display. **b)** A neutral learning trial with three distractors in the present study. Participants had to match the novel word presented both auditorily and visually onto one of the six pictures displayed on the screen. Note: In both illustrations we have replaced the IAPS pictures by other pictures to avoid copyright issues: a) A rattlesnake, from Pixabay.com. b) Pictures of our own: *(clockwise)* Luigi Santini, Tuscany, Aug. 2015; A fruit shop, China Town, Dec. 2016; Cows, Hainaut, Sept. 2016; Ostend harbour, July 2016.

Fig. 2. Sketch of the protocol, with its three distinct moments of 'encoding', 'immediate test' and 'consolidation test'.

Fig. 3. Latency (left) and correct response rate (right) for alarming and neutral trials, during the 20 blocks of the each encoding phase (i.e., Set 1 and Set 2 of picture-word associations). Top: for the 0-hr group. Bottom: for the 6-hr group. The increase in the number of distractors, from one to three (between Blocks 9 and 10), and from three to six (between Blocks 16 and 17) is indicated by the increased darkness of the grey-shaded area. Chance on the accuracy graphs (on the right) is indicated by the horizontal line, respectively at 50%, 75% and 83.5%.

Fig. 4. Picture-word association performance at the end of each encoding phase (i.e., Set 1 vs. Set 2), and at the combined test (i.e., Same Day vs. After 7 Days). **a)** Correct response rates for the whole sample (left), as well as for the 0-hr group (top right) and the 6-hr group (bottom right), separately. **b)** Latencies presented in a similar fashion. Error bars show standard errors. Also displayed are the *p*-values for simple effects and interactions of interest.

Fig. 5. a) Pause detection latencies from word onset (across pause-present and pause-absent trials), as a function of memory age and emotionality. Left: for the whole sample. Top right: for the 0-hr group only. Bottom right: for the 6-hr group only. **b)** Latencies in the emotional Stroop presented in a similar fashion. Error bars show standard errors. Also displayed are the *p*-values for simple effects and interactions of interest.

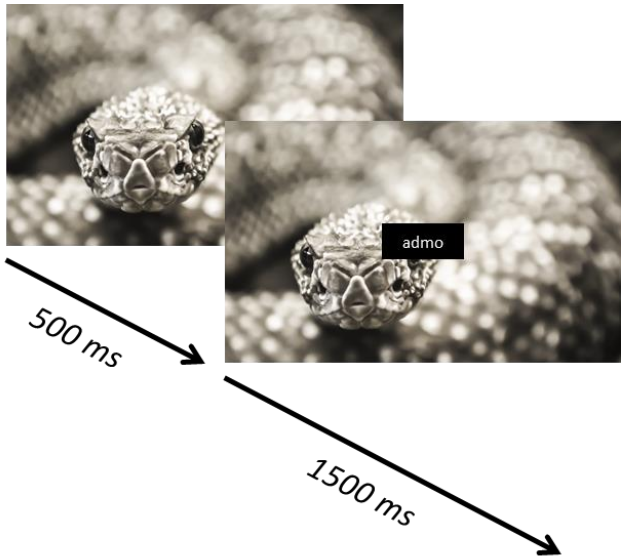
Fig. 6. Stroop latencies as a function of the age of trial *n* and of trial *n-1*. **a)** For trial *n*, overall (left) and in those 32 participants with the smallest response repetition priming (across all conditions,

except when both trials were seven-day old), separately for pairs of trials with unrepeated colour responses (top right) and pairs of trials with repeated colour responses (bottom right). **b)** For trial *n-1*, overall. Error bars show standard errors. Also displayed are the *p*-values for contrasts of interest.

Fig. 7. Stroop latencies on trial *n* as a function of the age of trial *n-1* and emotionality. **a)** When trial *n* taps into a seven-day-old (possibility consolidated) association. **b)** When trial *n* taps into a same-day (unconsolidated) association. Error bars show standard errors. Also displayed are the *p*-values for simple effects and interactions of interest. Note: The emotional Stroop is assessed one-tailed, as it only comes about as an interference (not a facilitation) effect.

Fig. 1

a)


admo
500 ms
1500 ms

b)

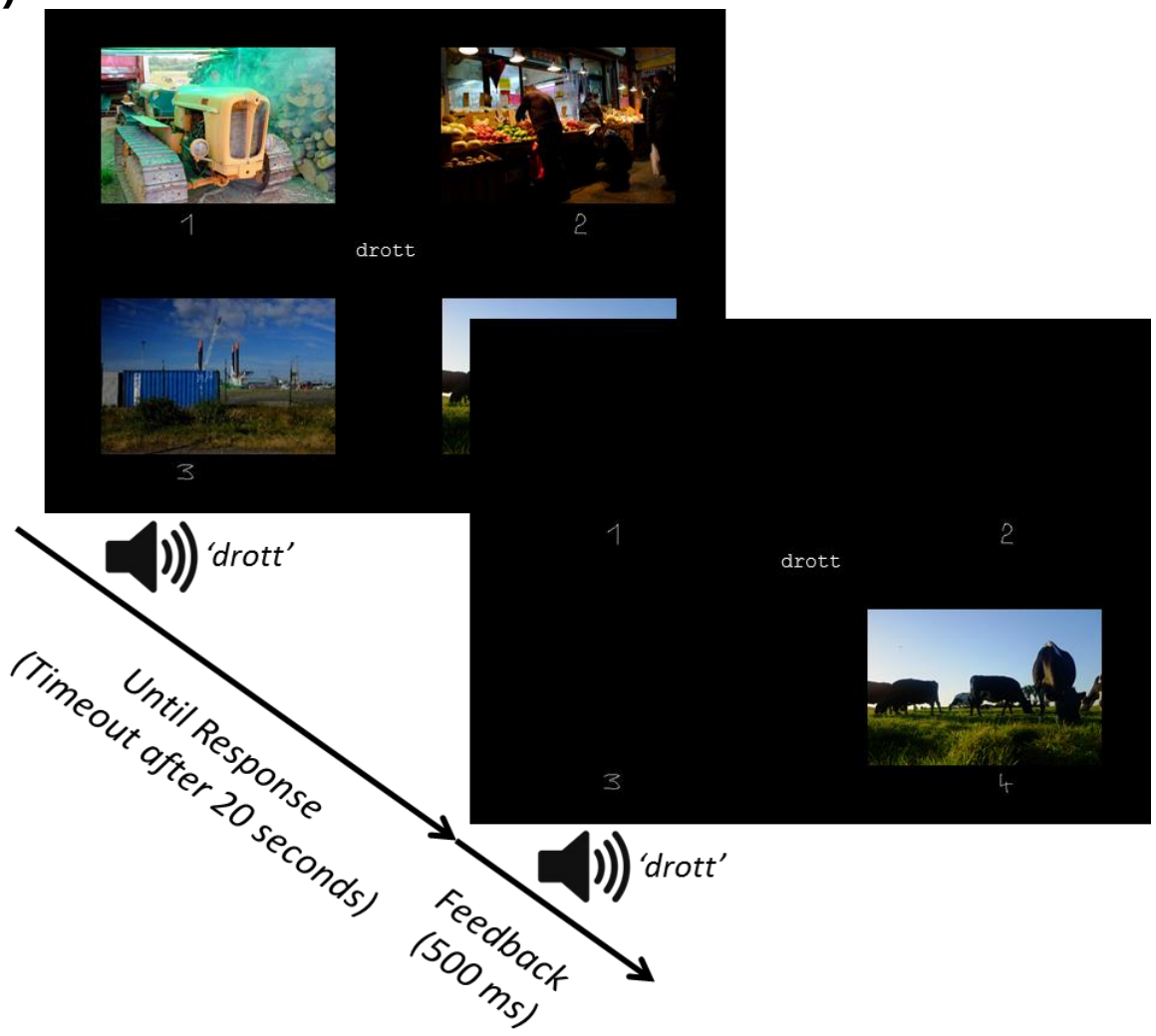

1
2
drott
3
'drott'
Until Response
(Timeout after 20 seconds)
1
2
drott
3
4
'drott'
Feedback
(500 ms)

Fig. 2

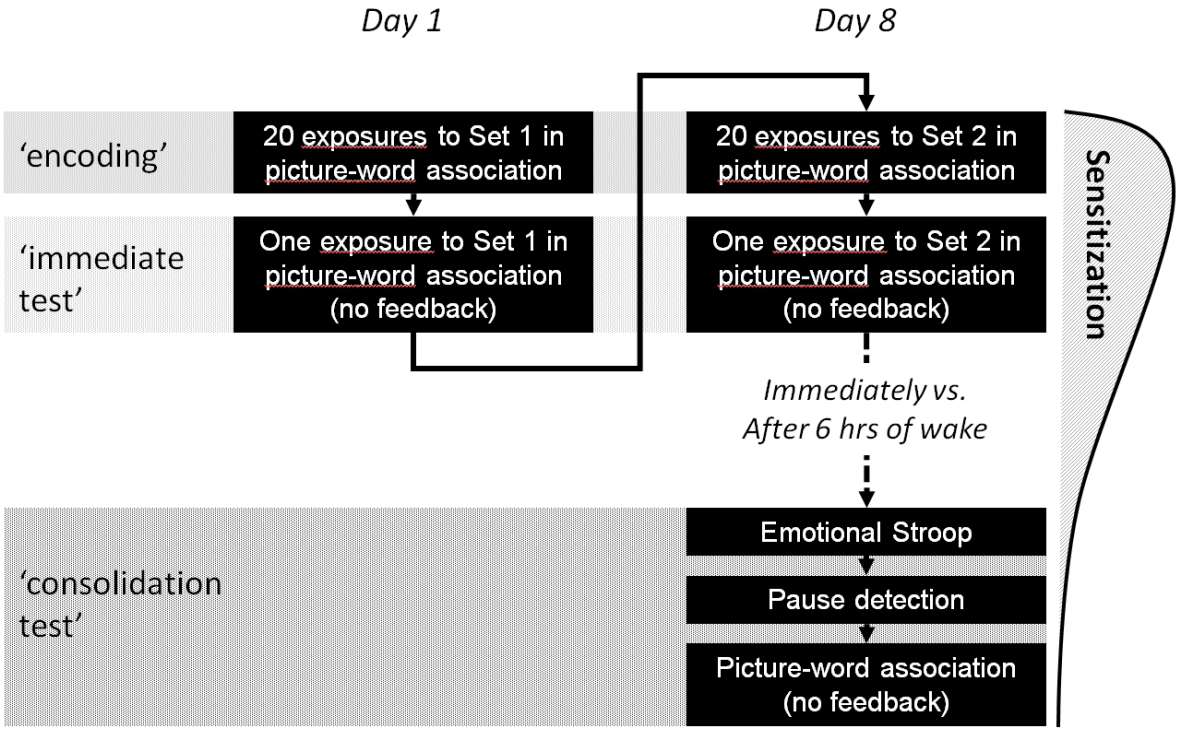

Day 1
Day 8
'encoding'
20 exposures to Set 1 in picture-word association
20 exposures to Set 2 in picture-word association
'immediate test'
One exposure to Set 1 in picture-word association (no feedback)
One exposure to Set 2 in picture-word association (no feedback)
Immediately vs. After 6 hrs of wake
'consolidation test'
Emotional Stroop
Pause detection
Picture-word association (no feedback)
Sensitization

Fig. 3

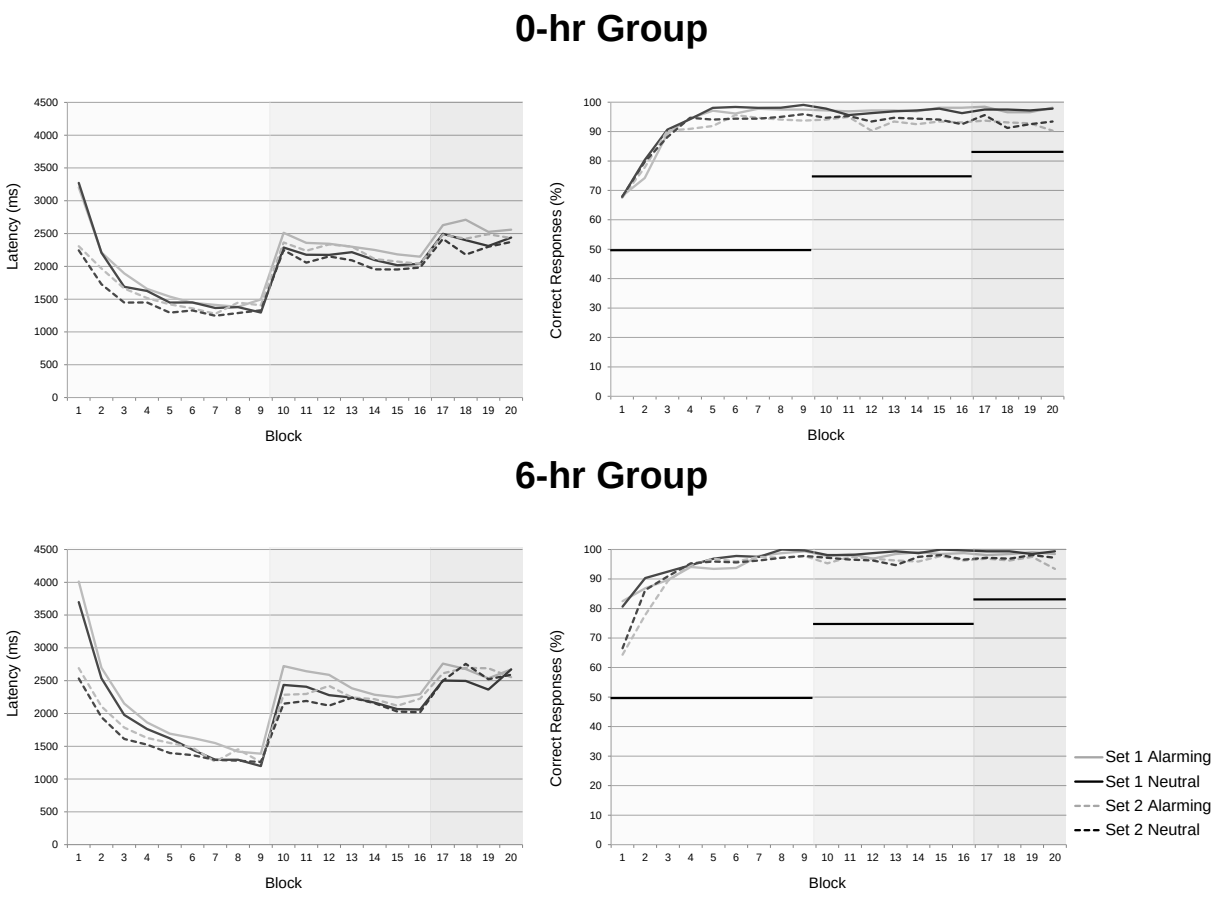

0-hr Group
Latency (ms)
Block
Correct Responses (%)
Block
6-hr Group
Latency (ms)
Block
Correct Responses (%)
Block
Set 1 Alarming
Set 1 Neutral
Set 2 Alarming
Set 2 Neutral

Fig. 4

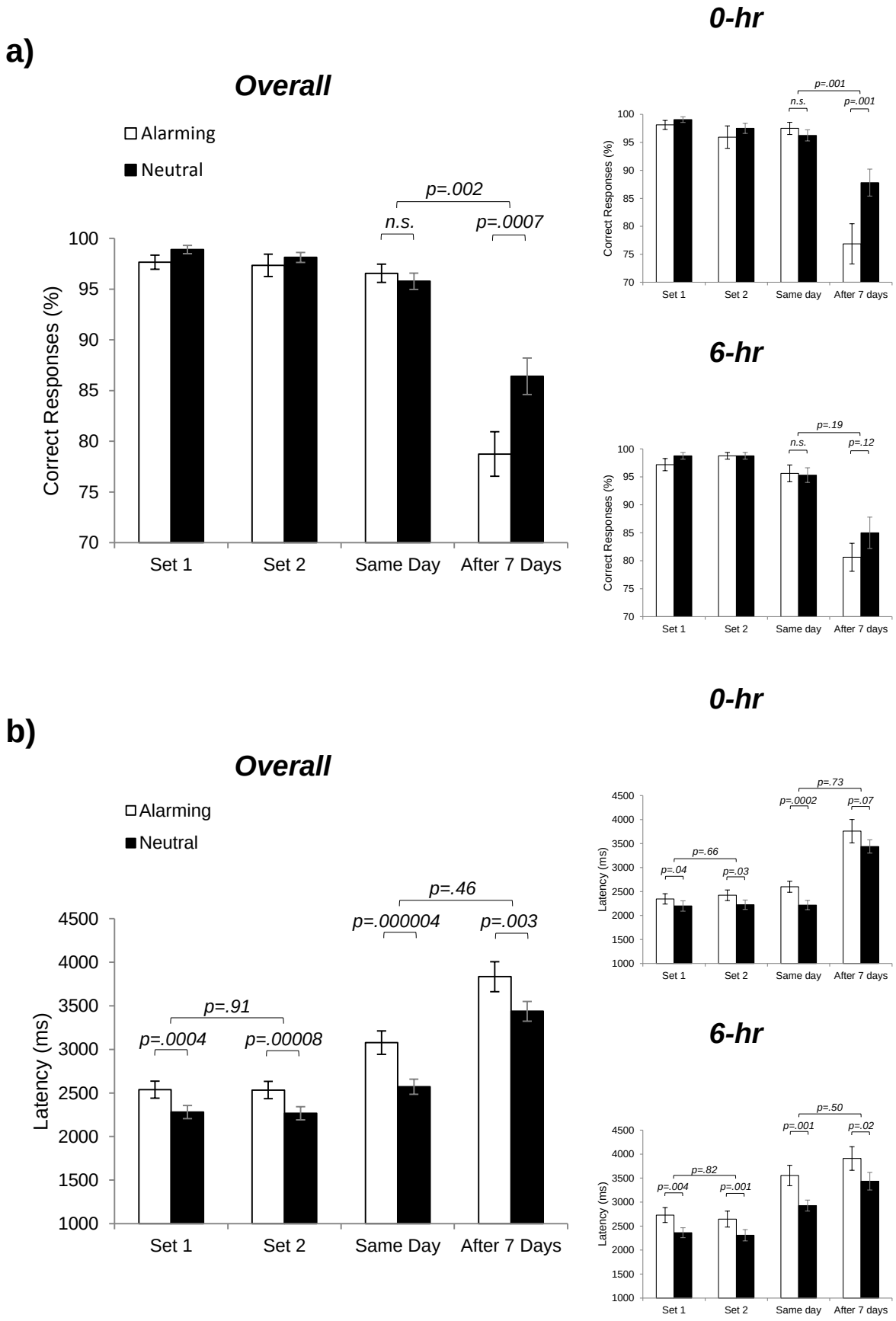
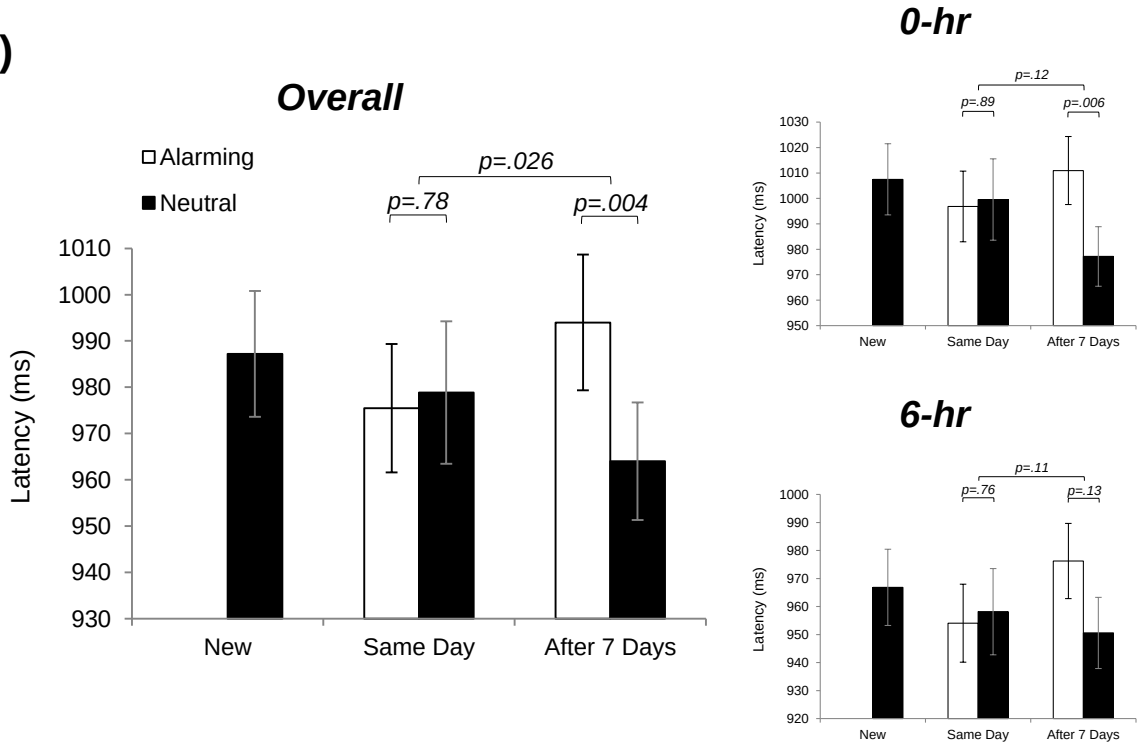

a)
Overall
Alarming
Neutral
p=.002
n.s.
p=.0007
Correct Responses (%)
100
95
90
85
80
75
70
Set 1
Set 2
Same Day
After 7 Days
0-hr
p=.001
n.s.
p=.001
Correct Responses (%)
Set 1
Set 2
Same day
After 7 days
6-hr
p=.19
n.s.
p=.12
Correct Responses (%)
Set 1
Set 2
Same day
After 7 days
b)
Overall
Alarming
Neutral
p=.46
p=.000004
p=.003
p=.91
p=.0004
p=.00008
Latency (ms)
4500
4000
3500
3000
2500
2000
1500
1000
Set 1
Set 2
Same Day
After 7 Days
0-hr
p=.73
p=.0002
p=.07
p=.66
p=.04
p=.03
Latency (ms)
Set 1
Set 2
Same day
After 7 days
6-hr
p=.50
p=.001
p=.02
p=.82
p=.004
p=.001
Latency (ms)
Set 1
Set 2
Same day
After 7 days

Fig. 5

a)


Overall
Alarming
Neutral
p=.026
p=.78
p=.004
Latency (ms)
1010
1000
990
980
970
960
950
940
930
New
Same Day
After 7 Days
0-hr
p=.12
p=.89
p=.006
1030
1020
1010
1000
990
980
970
960
950
Latency (ms)
New
Same Day
After 7 Days
6-hr
p=.11
p=.76
p=.13
1000
990
980
970
960
950
940
930
920
Latency (ms)
New
Same Day
After 7 Days

b)

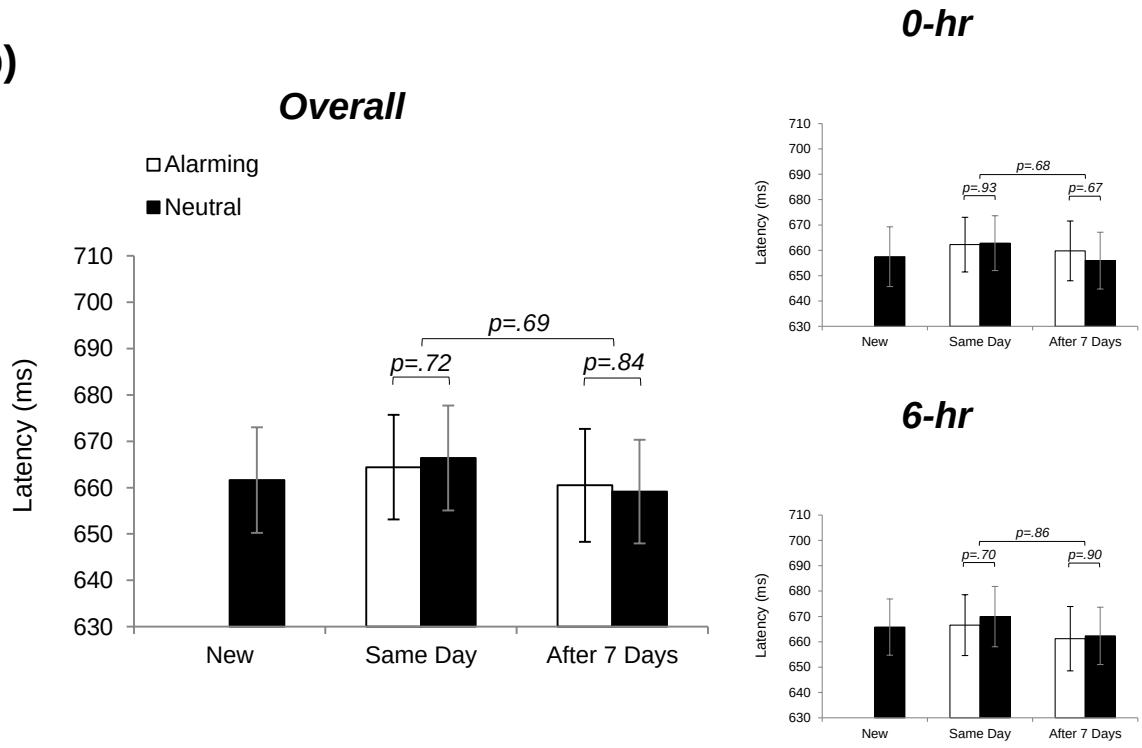
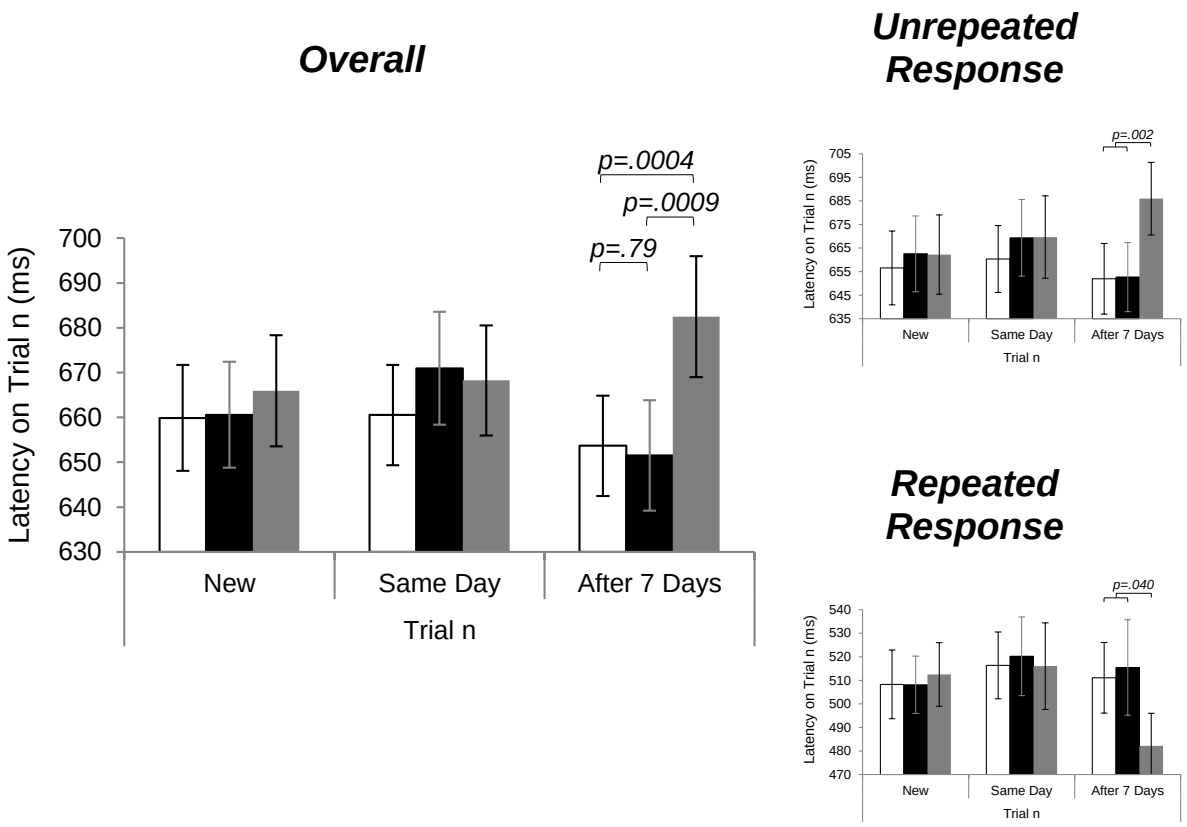

Overall
Alarming
Neutral
p=.69
p=.72
p=.84
Latency (ms)
710
700
690
680
670
660
650
640
630
New
Same Day
After 7 Days
0-hr
p=.68
p=.93
p=.67
710
700
690
680
670
660
650
640
630
Latency (ms)
New
Same Day
After 7 Days
6-hr
p=.86
p=.70
p=.90
710
700
690
680
670
660
650
640
630
Latency (ms)
New
Same Day
After 7 Days

Fig. 6

a)


Overall
Latency on Trial n (ms)
700
690
680
670
660
650
640
630
p=.0004
p=.0009
p=.79
New
Same Day
After 7 Days
Trial n
Unrepeated Response
Latency on Trial n (ms)
705
695
685
675
665
655
645
635
p=.002
New
Same Day
After 7 Days
Trial n
Repeated Response
Latency on Trial n (ms)
540
530
520
510
500
490
480
470
p=.040
New
Same Day
After 7 Days
Trial n

b)

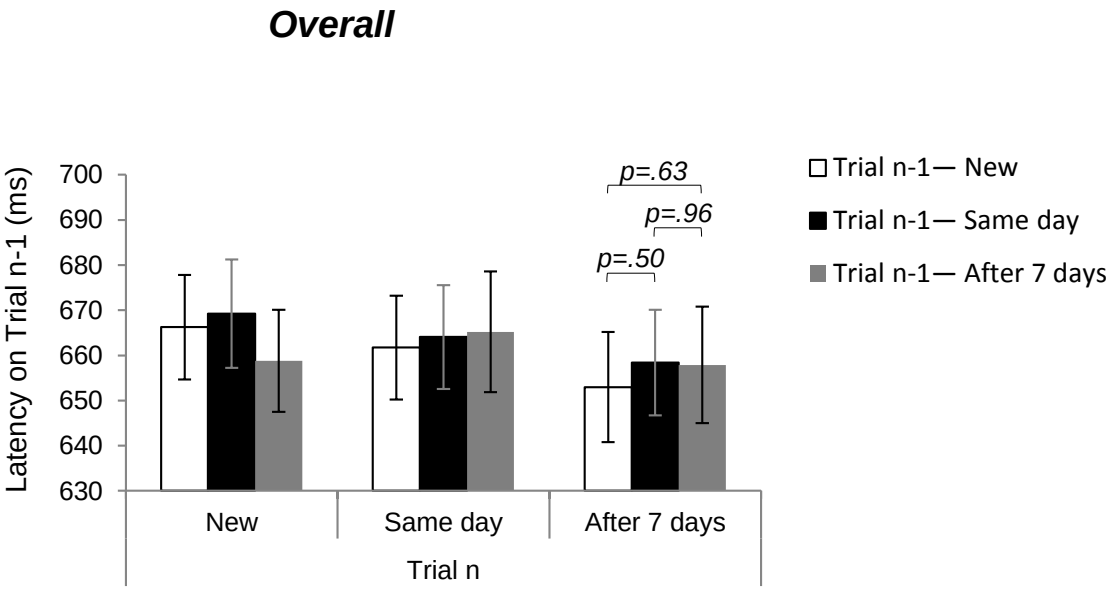
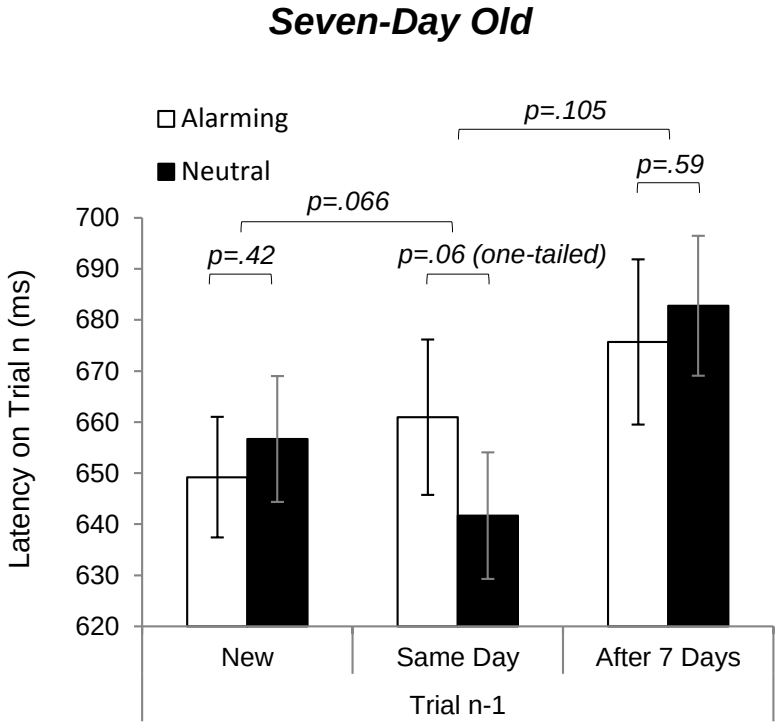
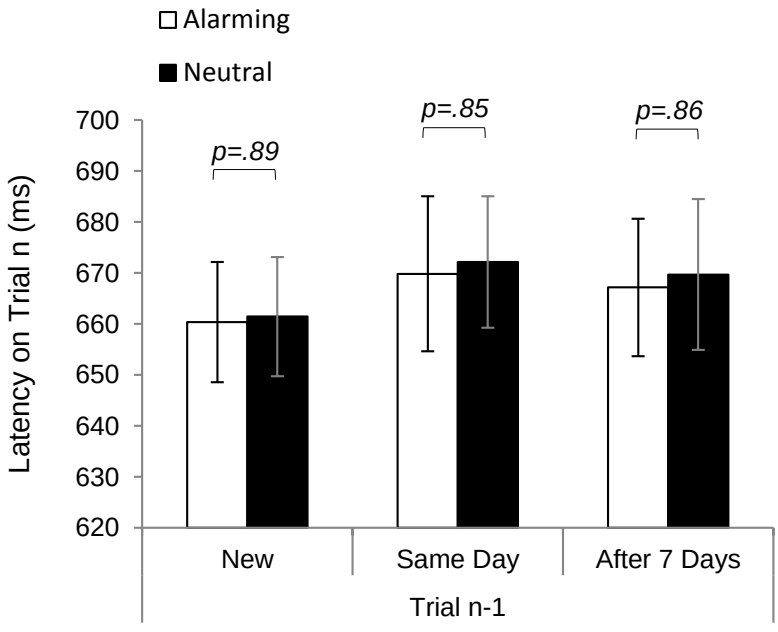

Overall
Latency on Trial n-1 (ms)
700
690
680
670
660
650
640
630
p=.63
p=.96
p=.50
New
Same day
After 7 days
Trial n
Trial n-1— New
Trial n-1— Same day
Trial n-1— After 7 days

Fig. 7

a)


Seven-Day Old
Alarming
Neutral
p=.105
p=.59
p=.066
p=.42
p=.06 (one-tailed)
Latency on Trial n (ms)
700
690
680
670
660
650
640
630
620
New
Same Day
After 7 Days
Trial n-1

b)


Same-Day
Alarming
Neutral
p=.85
p=.86
p=.89
Latency on Trial n (ms)
700
690
680
670
660
650
640
630
620
New
Same Day
After 7 Days
Trial n-1