

Can semantic relatedness explain the enhancement of memory for emotional words?

DEBORAH TALMI and MORRIS MOSCOVITCH
University of Toronto, Toronto, Ontario, Canada

Memory for emotional items is often better than memory for neutral items. In three experiments, we examined whether this typical finding is due to the higher semantic relatedness inherent to emotional items, a confound in previous studies. We also controlled for other possible confounding variables, such as imagery. In Experiments 1 and 2, participants encoded lists of emotional and categorized neutral words equivalent in semantic relatedness, as well as lists of random neutral words with lower semantic relatedness. In Experiment 3, the lists were mixed, containing words from all the conditions. Surprise free recall was tested after a 40- to 55-min retention interval. Free recall of emotional words was better than that of random neutral words, replicating the classic effect. Importantly, categorized words were recalled better than random neutral words, and not worse than emotional words. These results emphasize the important role of semantic relatedness in the classic effect and suggest that organizational processes operate alongside arousal-related ones to enhance memory for emotional material.

Emotional events are believed to be more memorable than neutral events. Research has shown that in free recall tasks, people remember emotionally evocative words (e.g., Colombel, 2000; Doerksen & Shimamura, 2001; Kensinger, Brierley, Medford, Growdon, & Corkin, 2002; LaBar & Phelps, 1998; Phelps, LaBar, & Spencer, 1997; Rubin & Friendly, 1986) and pictures (e.g., Blake, Varnhagen, & Parent, 2001; Bradley, Greenwald, Petry, & Lang, 1992; Buchanan, Denburg, Tranel, & Adolphs, 2001; Hamann, Cahill, & Squire, 1997; Hamann, Ely, Grafton, & Kilts, 1999; Kensinger et al., 2002; Palomba, Angrilli, & Mini, 1997) better than neutral stimuli. In the present study, we examined whether the effect of emotionality on memory is direct or mediated by other factors, such as semantic relatedness, known to influence memory.

In studies on the effect of emotion on memory, investigators manipulate arousal and valence of the material but rarely control for other factors that are known to influence memory. This problem can be especially serious when a variable is systematically confounded with emotionality. Semantic relatedness is such a confounding factor, because emotional stimuli are inherently more interrelated than unselected neutral stimuli. For instance, a picture of a dead body and a gun or the words *torture* and *misery* are strongly related; this is not the case for neutral stimuli, which are usually randomly selected pictures or words. Because interrelated stimuli are better remembered (e.g., Mandler, 1967; Puff, 1970; Tulving &

Pearlstone, 1966), it is necessary to determine to what extent semantic relatedness contributes to the enhancement of memory for emotional stimuli. If emotionality is a factor over and above semantic relatedness, emotional stimuli should be recalled better than equally related neutral stimuli. The present experiment tested this hypothesis.

We chose to use words in the present study, since it is easier to control words for factors related to memory than it is to control pictures, where the stimulus can be very rich. This is especially the case for photographs of emotional scenes, where emotionality may be correlated with unusualness or novelty (Adolphs, Denburg, & Tranel, 2001), interest (Hamann et al., 1999), and the scene's visual complexity (e.g., Ochsner, 2000). For similar reasons, we avoided using taboo words that form a closely related unit. The emotional words chosen in the experiments reported below were nontaboo, arousing, negatively valenced words and were rated as significantly different on the latter two dimensions, relative to the neutral words.

Another factor possibly correlated with the emotionality of words is imagery. It was necessary to control for imagery, since it has been found to be an even more important predictor of recall than emotionality (Rubin & Friendly, 1986). Rubin and Friendly found that for 925 nouns, arousing nouns were lower on imagery. When words are intentionally selected in a particular study to be extreme in arousal and valence, this correlation may or may not hold. Interestingly, our observations indicate that semantic relatedness is positively correlated with imagery for neutral words. For example, it is apparent that published lists of categorized neutral words (e.g., Battig & Montague, 1969) are higher on imagery than are lists of emotional words. In order to compare emotional and categorized neutral words properly, one should also control for imagery; otherwise, emotional words are

The authors thank Eyal Reingold and Ulrich Schimmack for helpful comments and Ming Lee for help in running Experiment 2. Correspondence concerning this article should be addressed to D. Talmi or M. Moscovitch, Department of Psychology, University of Toronto, 100 St. George Street, Toronto, ON, M5S 3G3 Canada (e-mail: debbie@psych.utoronto.ca or momos@psych.utoronto.ca).

placed at a disadvantage. On the other hand, taboo words are considered highly emotional, but also high on imagery, which may partly account for their high memorability. Imagery and emotionality, therefore, might be confounded, or alternatively, this variable might add only to random error of measurement within a study or to the noise in the overall pattern of findings. To overcome these possibilities and the difficulties they cause in determining the effect of emotion on memory, we controlled for imagery of the words, as well as for familiarity, frequency, and word length.

To examine the effects of relatedness on memory, we compared memory for emotional, random neutral, and categorized neutral words. Experiments 1 and 2 used pure lists of words from each condition, whereas Experiment 3 extended the findings to a mixed list design. Three different incidental encoding tasks were used across experiments: reading aloud in Experiment 1, silent reading in Experiment 2A, and familiarity rating in Experiments 2B and 3. The categorized neutral words belonged to two categories (kitchen and music) in Experiment 1 and to a third category (driving) in Experiments 2 and 3. All the experiments used a distractor task to fill the interval between study and test. The length of the distractor task (40–45 min in Experiment 1 and 50–55 min in Experiments 2 and 3) was modeled after Kleinsmith and Kaplan's (1963) finding that a delay longer than 20 min is sufficient to allow a memory advantage for emotional material to surface. The same delay has been used successfully by other researchers with verbal stimuli (e.g., Kensinger et al., 2002; LaBar & Phelps, 1998). As we will demonstrate below, we replicated the classic finding, showing enhanced memory performance for emotional, relative to random neutral, words. Importantly, memory for categorized neutral words was as good as or better than memory for emotional words.

EXPERIMENT 1

In this first experiment, each participant studied a single list of words in an incidental encoding procedure, was then engaged in a distractor task, and was subsequently asked to recall freely the previously studied words. We used incidental encoding and an attention-demanding nonverbal distractor task in order to minimize potential differences at encoding and rehearsal. These considerations are important because, under intentional encoding, participants might use differential encoding strategies or differential rehearsal patterns. We aimed to avoid those differences, which might confound the memory effect (but see Guy & Cahill, 1999).

We chose a shallow reading task rather than the more typical emotionality rating task used in previous research (e.g., LaBar & Phelps, 1998). The explicit use of the emotionality dimension at encoding in previous studies possibly contributed to the memory enhancement for emotional words, because it could serve as an externally provided organizational framework at encoding or be used as a retrieval cue.

Method

Participants. Undergraduate students at the University of Toronto received course credit or were paid \$10 for their participation. Eleven participants were replaced, 9 because they were classified as *aware* (see below) and 2 due to experimenter error. Final data were collected from 60 participants (19 males, 41 females; mean age, 20 years, $SD = 3.7$). In this and in the following experiments, all the participants were native English speakers or had received all their education (Grade 1 and up) in English. In accordance with ethics committee guidelines, the participants signed a consent form before starting and were thanked and debriefed when finished.

Materials. Six 28-word lists were used in this experiment: two emotionally arousing, negatively valenced lists (hereafter, *emotional lists*), two categorized neutral lists, and two random neutral lists. Two sets were prepared, each containing one list of each type. Each participant was randomly allocated to study only one list. The means of the six lists were not significantly different from one another in familiarity ($M = 526.77$, $SD = 56.9$), imagery ($M = 510.68$, $SD = 74.75$), frequency ($M = 48.10$, $SD = 61.15$), and number of letters ($M = 5.74$, $SD = 4.66$). Values for these variables were obtained from published norms (Coltheart, 1981). A frequency value of zero was assigned to 2 words with no Kučera–Francis value (*oboe* and *suffocate*; related words, *oboist* and *suffocation*, had a Kučera–Francis value of 1). A multivariate analysis of variance (ANOVA) showed that all six lists were not significantly different from each other on any of the above variables ($p > .10$). Six buffer words per list, 3 at the beginning and 3 at the end, were sampled from an additional list of 24 neutral words to minimize primacy and recency effects.

The emotional and the categorized word lists were matched for semantic relatedness on the basis of an extensive series of pilot studies. Two separate groups of 6 participants each rated the emotional and the categorized neutral lists from the same set, and a third group of 5 participants rated the two emotional and the two random neutral lists (3 participants who correlated below .4 with other raters were replaced). The participants were presented with all possible pairings of each of the words in a particular list and were asked to judge how much the two words in each pair were related on a scale of 1 (*low association*) to 7 (*high association*). Presentation order of list pairs and word presentation within its pairs (left or right) were random. Each word in the list received a relatedness score computed as the average score of its 27 pairs. These scores were used to test interrater reliability, which was found to be sufficiently high (standardized item alpha was .88 for Group 1, .85 for Group 2, and .94 for Group 3). A list relatedness score was computed as the average score of all list words for each participant. For the first two groups, a repeated measures ANOVA with list type as a within-subjects factor and set as a between-subjects factor showed no significant differences in semantic relatedness between the emotional and the categorized lists [$F(1,1) = 0.34$, $p > .10$] and no other significant differences. For Group 3, a repeated measures ANOVA with list type and set as within-subjects factors showed that the emotional lists were significantly more cohesive than the random lists [$F(1,4) = 26.49$, $p < .01$]. There was also a type \times set interaction, so that the differences were stronger in Set 2 [$F(1,4) = 8.33$, $p < .05$].

Finally, a separate group of 18 participants judged a randomly mixed list of all the words used in this study, as well as additional words (total, 331), for valence and arousal. One participant was replaced due to low correlation (less than .1) with all other participants. Interrater reliability was high (standardized item alpha for valence was .98 and, for arousal, .94). For each set, two separate repeated measures ANOVAs were conducted on the valence and the arousal scores for the three list types (see Table 1). The absolute values of the valence scores for emotional words were used in the analysis to provide a more conservative estimate of the differences between the emotional and the neutral lists. The effect of list type was significant for both valence [Set 1, $F(2,34) = 197.50$, $p < .001$; Set 2, $F(2,34) = 220.73$, $p < .001$] and arousal [Set 1, $F(2,34) = 120.07$, $p < .001$; Set 2, $F(2,34) = 360.49$, $p < .001$]. Planned com-

Table 1
Valence and Arousal Ratings of Experimental Word Lists

	List Type					
	Emotional		Categorized		Random Neutral	
	Set/List* 1	Set/List 2	Set/List 1	Set/List 2	Set/List 1	Set/List 2
Experiment 1						
Valence						
<i>M</i>	-1.8	-1.98	0.067	0.87	0.20	0.20
<i>SD</i>	0.26	0.25	0.37	0.43	0.18	0.19
Arousal						
<i>M</i>	3.78	3.85	1.49	1.65	0.70	0.66
<i>SD</i>	0.60	0.67	1.04	1.01	0.55	0.61
Experiment 2						
Valence						
<i>M</i>	-1.79		0.30		0.32	
<i>SD</i>	0.26		0.21		0.24	
Arousal						
<i>M</i>	3.68		0.92		1.19	
<i>SD</i>	0.60		0.82		0.69	
Experiment 3						
Valence						
<i>M</i>	-2.06	-2.40	0.38	0.28	0.33	0.26
<i>SD</i>	0.37	0.27	0.21	0.34	0.29	0.24
Arousal						
<i>M</i>	4.03	4.55	1.09	0.90	1.14	0.65
<i>SD</i>	0.91	0.84	0.81	0.94	0.81	0.58

Note—Valence values refer to participants' ratings on a -3 to +3 scale (*very negative* to *very positive*). Arousal values refer to participants' ratings on a 0-6 scale (*not arousing at all* to *very arousing*). *The numbers refer to the sets in Experiment 1 and to the lists in Experiment 3.

parisons showed that in accord with our expectation, the emotional words had more extreme valence ($p < .001$), as compared with the random and the categorized neutral words. In addition, contrary to our expectation, the categorized neutral words had slightly higher valence and arousal ($p < .001$) than did the random neutral words.

The words were presented centrally in 24-point black Times New Roman font on a white background. The words were presented with Authorware Professional 6.0 software. Nonverbal nonmnemonic distractor tasks were used in this study and were presented with E-Prime software.

Procedure. The participants were tested individually in a sound-attenuated room. The experiment was introduced as a study of attention. Accordingly, the distractor tasks that were used to fill the retention interval were computer-based tasks of speeded visual and auditory attention. The encoding task was described as simply the first attention task, examining attention in reading, and the participants' reading was taped. This was done to minimize the risk that the participants would guess that their memory would be tested later. The words were presented serially at a rate of 4 sec/word. The participants were asked to attend to the words and read them out loud as they appeared.

At the end of the encoding task, the experimenter instructed the participants on how to proceed with the distractor tasks on their own. The retention interval lasted between 40 and 45 min. When finished, the participants were asked the following question: "What did you think was the purpose of the first task, when you read words and I recorded your reading?" The participants who indicated that they had expected a memory test were classified as *aware*, and their data were discarded. The participants were then given 3 min to record all the words they remembered on a blank sheet of paper.

Results

In this and all the other experiments reported below, variant spellings of experimental words were considered

correct (e.g., *depression* for *depressed*); using a stringent criterion did not change the results. One of the random neutral lists inadvertently contained two pairs of words that rhymed with each other (*manner/banner*; *tool/stool*). The problem that ensued was that this unfortunate connection helped some participants, but not others, leading to a high variance in this group and a non-homogeneous ANOVA. Our solution to the problem was to eliminate those words from the analysis and score the participants' responses out of 24 for that list, rather than out of 28.

Percentage of recall scores were analyzed with a 2 (set) \times 3 (type) ANOVA. There was a significant effect of list type on the percentage of free-recalled words [$F(2,54) = 25.14, p < .001$; see Figure 1]. The effect size of list type was large (Cohen's $d = 1.93$). Planned comparisons showed that free recall of emotional words ($p < .001$) and categorized words ($p < .001$) was better than that of random neutral words, replicating the classic effects. Importantly, post hoc tests showed that categorized neutral words were recalled slightly, but not significantly, better than emotional words ($p > .10$). None of the effects of set was significant.

Discussion

Negatively valenced, arousing words were recalled better than random neutral words, replicating the classic finding. Importantly, free recall of categorized neutral words was equivalent to free recall of emotional words, but higher than free recall of random neutral words. The

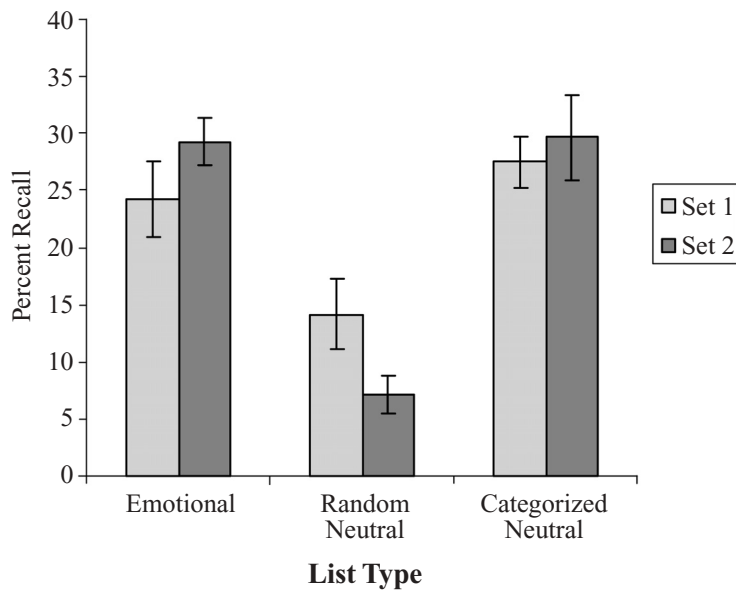


Figure 1. Free recall performance (in percentages) as a function of list type in Experiment 1.

improvement in free recall of categorized words reflects their higher semantic relatedness, a well-known effect in the memory literature (e.g., Mandler, 1967; Puff, 1970; Tulving & Pearlstone, 1966). The most parsimonious interpretation of the enhanced free recall for emotional words in this experiment is their increased interrelatedness relative to that for random words, rather than to emotionality per se. Had emotionality played a role in improving memory for these words, it should have increased recall beyond that observed for categorized words.

A possible weakness of Experiment 1 was that the categorized neutral words were found in pilot testing to be more arousing and of greater valence than the random neutral words. Although the difference between the random and categorized neutral lists was nominally small, and much smaller than the difference in emotionality between both of them and the emotional lists, it is still possible that the higher recall of the categorized neutral words stemmed from their increased emotionality. To rule out this alternative, we attempted to replicate the findings of Experiment 1 with an improved stimulus set.

EXPERIMENT 2A

Experiments 2A and 2B replicated and extended the findings from Experiment 1, using different encoding tasks (silent reading in Experiment 2A and familiarity rating in Experiment 2B) and a new category for the categorized neutral condition. With hindsight, we realized that the reason for the higher emotionality of the categorized neutral list in Experiment 1 was the choice of the generally positive categories of kitchen and music. In Experiment 2, we chose driving-related words.

The stimulus set created for Experiment 2 consisted of three lists: an emotional list, a categorized neutral list,

and a random neutral list. This time, however, the categorized and the random neutral lists were better matched on valence and arousal. As in Experiment 1, each participant studied only one list of words.

Method

Participants. Undergraduate students at the University of Toronto were paid \$10 for their participation. Four participants were classified as *aware* and were replaced. Final data were collected from 54 participants (16 males, 38 females; mean age, 20.277 years, $SD = 4.448$).

Materials. A random list of neutral words and a categorized list of driving-related words were constructed, matched with each other for valence and arousal. Each participant was randomly allocated to study only one of these lists. These neutral lists and one of the emotional lists from Experiment 1 were also matched for familiarity ($M = 535.56$, $SD = 43.533$), imagery ($M = 497.29$, $SD = 76.04$), frequency ($M = 69.62$, $SD = 112.45$), and number of letters ($M = 5.83$, $SD = 2.02$), so that their means did not differ significantly ($p > .10$). The effect of list type was significant for absolute valence [$F(2,34) = 270.29$, $p < .001$] and arousal [$F(2,34) = 132.81$, $p < .001$]. The emotional list had higher absolute valence and higher arousal than the categorized and the random lists ($p < .001$). The categorized neutral list and the random neutral list did not differ significantly in valence ($p > .10$), but the categorized neutral list was lower in arousal ($p < .05$; see Table 1).

The three lists were given, in a random order, to a separate group of 12 participants, who judged them for relatedness, following the procedure described in Experiment 1 (standardized item $\alpha = .96$). A repeated measures analysis showed the effect of list type to be significant [$F(2,54) = 139.51$, $p < .001$]. The emotional and the categorized neutral lists had similar relatedness scores ($p > .10$), which were higher than the relatedness score of the random neutral list ($p < .001$).

The distractor tasks used in Experiment 1 were used here as well. **Procedure.** The procedure in Experiment 2A was identical to that in Experiment 1 in all aspects except for the following. In Experiment 2, the participants were tested in larger groups. Group testing required a change in the encoding task from reading aloud

to silent reading. As in Experiment 1, each participant was randomly allocated to study only one of the word lists, and each word was presented for 4 sec, and then replaced by the next word. Here, however, the participants pressed the space bar as soon as they finished mouthing each word silently.

Results

There was a significant effect of list type on the percentage of free-recalled words [$F(2,51) = 5.00, p = .01$; see Figure 2]. The effect size of list type was large (Cohen's $d = 0.88$). Planned contrasts showed that free recall of emotional words was better than that of random neutral words ($p < .05$), replicating the classic effect. Importantly, categorized words were recalled slightly, but not significantly, better than emotional words ($p > .10$). Post hoc tests showed that categorized neutral words were recalled significantly better than random neutral words ($p < .01$).

We were interested in examining the relative contribution of our variables to memory. To determine the effects of emotional arousal, relatedness, and imagery on memory, we conducted a simultaneous multiple linear regression analysis of the free recall scores for each word. We included arousal, rather than valence, because the predicted relationship of valence and recall is not linear and because arousal and absolute valence are highly correlated, resulting in collinearity problems ($r = .93, p < .001$). This analysis was not possible for Experiment 1, because of nonnormal distribution of the residuals. The overall model explained .341 of the variance [$F(3,80) = 13.81, p < .001$]. Relatedness was a significant predictor of free recall ($\beta = .54; t = 5.14, p < .001$).

Imagery was a marginally significant predictor of free recall ($\beta = .18; t = 1.87, p = .065$). No other predictor was significant ($p > .10$).

Discussion

These results replicated the previous findings with an improved stimulus set. We examined memory for three lists of words: an emotional list, a random neutral list, and a list of categorized neutral words matched for relatedness with the emotional list. As before, free recall of the emotional words was higher than free recall of random neutral words. Importantly, free recall of the categorized neutral words was higher than free recall of the random neutral words and equivalent to that of the emotional words. The same pattern was evident in the regression analysis, where relatedness and, marginally, imagery predicted memory performance, but emotionality did not.

EXPERIMENT 2B

Memory performance in Experiment 2A was lower than memory performance in Experiment 1. The cause for the decline in memory was possibly the shallow encoding task and, perhaps, the group-testing situation. Another potential problem with the silent-reading encoding task was that it involved a relatively long presentation of each word with little activity, possibly allowing the participants to employ differential encoding strategies in the three conditions. To improve memory performance within a group setting and avoid potential differences in

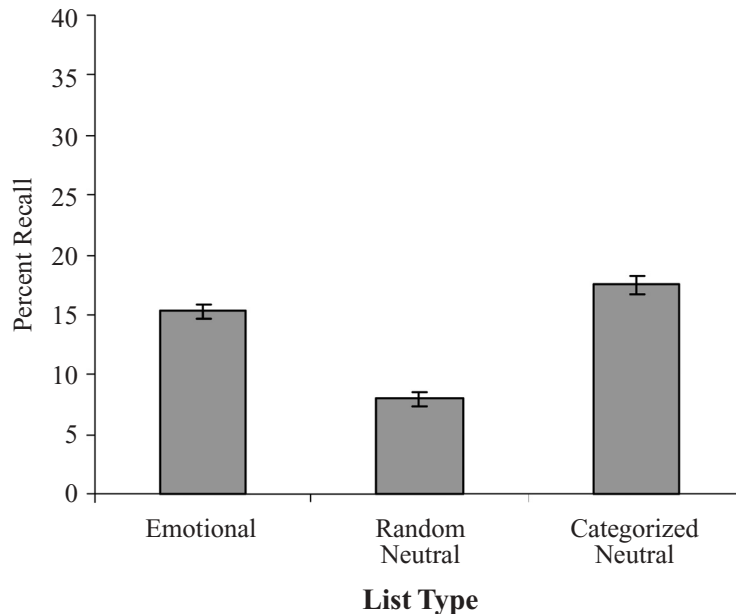


Figure 2. Free recall performance (in percentages) as a function of list type in Experiment 2A.

encoding strategies, Experiment 2B employed a fast-paced but deep familiarity rating encoding task. In all other respects, Experiments 2A and 2B were identical.

Method

Participants. Students at the University of Toronto were paid \$10 for their participation. Two participants were replaced, 1 because she was classified as *aware* and 1 due to experimenter error. Final data were collected from 54 participants (18 males, 36 females; mean age 24.27 years, $SD = 3.06$).

Materials and Procedure. The materials and procedure in Experiment 2B were identical to those in Experiment 2A in all respects except the following. The familiarity rating encoding task followed the procedure used by Toglia and Battig (1978). The participants were instructed to rate each presented word on a scale ranging from 1 (*low familiarity*, words never seen, heard, or used) to 7 (*high familiarity*, words seen, heard, or used every day). Also as in Toglia and Battig, the participants were told that since words differ in many ways, they were to make sure that they referred only to familiarity in their rating. Each word was presented for 1.5 sec and was then replaced by a row of *x*s. A rating scale was presented at the bottom of the screen, and the participants clicked with the mouse on the number that corresponded to their level of familiarity with the preceding word. A blank screen for 1,250 msec and an asterisk for 250 msec were then presented, followed by the next word. The retention interval was approximately 45 min.

Results

There was a significant effect of list type on the percentage of free-recalled words [$F(2,51) = 28.66, p < .001$; see Figure 3]. The effect size of list type was large (Cohen's $d = 2.12$). Planned contrasts showed that free recall of emotional words was better than that of random neutral words ($p < .01$), replicating the classic effect.

Importantly, the categorized words were recalled even better than the emotional words ($p < .001$).

We conducted a simultaneous multiple linear regression analysis of the free recall scores for each word with the predictors imagery, arousal, and relatedness. The overall model explained .314 of the variance [$F(3,80) = 12.20, p < .001$]. Relatedness was a significant predictor of free recall ($\beta = .62; t = 5.80, p < .001$). Arousal had a suppressor effect on free recall: It did not correlate with recall ($r = .01, p > .10$) but still served as a significant predictor ($\beta = -.25; t = -2.37, p < .05$). This effect remained when imagery was excluded from the model and was probably due to the positive correlation between emotional arousal and relatedness ($r = .40, p < .001$). This effect across words reflects our findings across participants. Since the emotional and the categorized neutral words were equal on semantic relatedness and the emotional words were recalled less well than the categorized neutral words, arousal effectively pushed memory down for those words.

The effect of type on familiarity rating and rating latency at encoding were examined in two separate univariate ANOVAs, yielding no significant results ($p > .10$).

The familiarity judgment task encouraged participants to relate the words to their own experiences with the words. A possible interpretation for the increased performance on categorized neutral items, relative to emotional items, in this experiment is that it was easier for participants in this condition to realize that the words in the list were related to each other. In order to examine

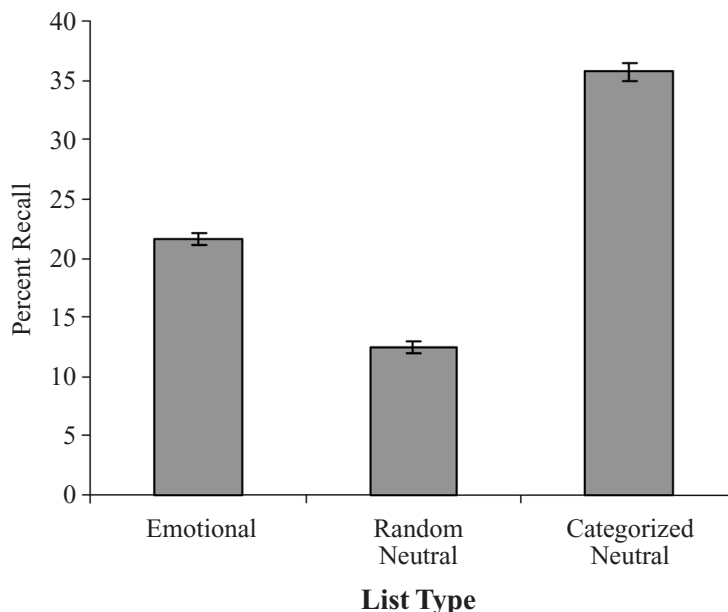


Figure 3. Free recall performance (in percentages) as a function of list type in Experiment 2B.

this possibility, we conducted phone interviews with 8 participants in the categorized neutral and 8 participants in the emotional conditions. The participants were first asked whether they had noticed anything special about the words at encoding, and then whether they had used any special strategy to try to recall the words at retrieval. The participants were then told what the nature of the list was and were asked whether they had noticed it at encoding or had used it at retrieval. Whereas only 2 participants in the emotional condition gave a positive answer to any of those questions, 6 participants in the categorized neutral condition gave a positive answer. This informal analysis indicated that it may have been easier for the participants in the categorized neutral condition to access the categorized nature of the list they studied.

Discussion

Experiment 2B replicated our previous results and showed that emotionality does not increase memory for words beyond the effect of semantic relatedness. In Experiments 1 and 2A, memory was equivalent for emotional and categorized neutral lists. In Experiment 2B, memory for the categorized neutral list was even higher than that for an equally cohesive emotional list, possibly because the deep encoding task focused the participants' attention on the semantic aspects of the words. As in the previous experiment, relatedness was a significant predictor of free recall.

EXPERIMENT 3

One difference between the previous experiments and most other investigations of emotional verbal memory is that whereas we used pure lists of words, previous studies have typically used mixed lists of neutral and emotional words (e.g., Colombel, 2000; Doerksen & Shimamura, 2001; Kensinger et al., 2002; Phelps et al., 1998). Having shown that emotional words are recalled better than random neutral words in pure lists, we attempted to extend the finding to a mixed list design. We prepared two new mixed lists. Each list consisted of words of the three word types used in the previous experiments (emotional, categorized neutral, and random neutral words).

In a mixed list design, emotional items might be recalled better not only because of their emotionality per se, but also because emotionality makes these items more distinct. Emotional items are considered more distinct than neutral items, an effect attributed to the unique physiological responses they evoke in the observer (e.g., Ochsner, 2000). Typically, distinctive items have an advantage in a mixed, but not a pure, list design (see McDaniel, DeLosh, & Merritt, 2000, for a brief review). The advantage of emotional over neutral items should then be larger in a mixed list design. This effect was demonstrated by Dewhurst and Parry (2000), where emotional words were remembered better than neutral words only in a mixed list design.

In contrast to Dewhurst and Parry (2000), we did obtain better memory for emotional words, relative to ran-

dom neutral words, in a pure list design in Experiments 1 and 2. We suggest that the difference between their study and ours is due to the different retrieval tasks used. Dewhurst and Parry used remember-know recognition judgments, whereas we used a free recall task. These two types of memory tests are known to probe memory for different kinds of information (e.g., Kinsbourne & George, 1974). In particular, organizational strategies—uniquely supporting memory for related items—play a larger role in free recall than in recognition.

Increased distinctiveness could lead to superior memory for emotional items even when relatedness is controlled. The following experiment tested the possibility that in a mixed list design, the emotional enhancement of memory would resurface.

Method

Participants. Undergraduate students at the University of Toronto were given course credit for their participation. Three participants were replaced, 1 because of an abnormal recall pattern (she recalled 14 words that were not on the study list, but only 4 studied words) and 2 due to experimenter error. Final data were collected from 40 participants (10 males, 15 females; mean age, 19.85 years, $SD = 4.16$).

Materials and Procedure. Two new lists were prepared. Each list had 27 words: 9 emotional words, 9 categorized neutral words, and 9 random neutral words. Twelve undergraduate students at the University of Toronto were paid \$10 to rate the relatedness of the words in the two lists (4 males, 8 females; mean age, 21.08 years, $SD = 2.06$), which were presented in a counterbalanced order, following the procedure described in Experiment 1. Three additional words from each condition were included in each list but were later discarded. A 2 (list) \times 2 (type) repeated measures ANOVA showed that the effect of type was significant [$F(2,22) = 213.39, p < .001$], as well as the set \times type interaction [$F(2,22) = 20.227, p < .001$]. To examine the effect of type for each list, we ran two separate univariate ANOVAs. In both, the effect of type was significant [List 1, $F(2,22) = 60.67, p < .001$; List 2, $F(2,22) = 258.53, p < .001$]. Planned contrasts showed that for both lists, the categorized neutral words were more interrelated than the random neutral words [List 1, $F(1,11) = 126.98, p < .001$; List 2, $F(1,11) = 293.89, p < .001$], but the difference between the categorized neutral words and the emotional words was not significant ($p > .10$).

For both lists, a multivariate ANOVA revealed that the means of the three word types were not significantly different from one another in familiarity, imagery, frequency, and number of letters (all $p > .10$; see Table 2). The effect of list type was significant for absolute valence [List 1, $F(2,34) = 217.93, p < .001$; List 2, $F(2,34) = 400.38, p < .001$] and for arousal [List 1, $F(2,34) = 85.85, p < .001$; List 2, $F(2,34) = 119.51, p < .001$]. Planned contrasts revealed that for both lists, the categorized and the random neutral words were not significantly different in absolute valence ($p > .10$) or arousal ($p > .10$). The emotional words, however, had higher absolute valence and arousal scores, relative to the neutral words ($p < .001$; see Table 1).

We used a different distractor task in this experiment, which was also nonverbal and nonmnemonic. The retention interval lasted approximately 55 min. The rest of the procedure was identical to the one employed in Experiment 2B.

Results

We examined the effect of word type on the percentage of free-recalled words with a 2 (list) \times 3 (word type) repeated measures ANOVA. The effect of word type was significant [$F(2,76) = 36.62, p < .001$; see Figure 4].

Table 2
Familiarity, Imagery, Frequency, and Word Length of Words in Experiment 3

Word Type	Familiarity		Imagery		Frequency		Word Length	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
List 1								
Emotional	553.22	36.14	494.11	70.32	79.11	82.81	4.78	1.30
Random	522.56	40.42	485.89	81.54	60.67	54.42	6.33	1.58
Categorized	536.44	51.61	542.44	61.82	61.78	81.10	5.44	2.19
List 2								
Emotional	515.89	66.95	517.78	47.40	84.22	147.14	5.67	2.18
Random	521.11	61.36	560.22	30.47	24.89	23.85	5.78	1.64
Categorized	546.00	55.22	552.67	67.26	59.56	58.89	5.89	2.15

The effect size of word type was large (Cohen's $d = 1.96$). Planned contrasts showed that free recall of emotional words was better than that of random neutral words ($p < .001$), replicating the classic effect. Importantly, the categorized words were recalled marginally better than the emotional words ($p = .061$). Post hoc tests confirmed that the categorized words were recalled better than the random neutral words ($p < .01$). There were no other significant effects.

The effect of type on latency to rate the words at encoding was examined in a repeated measures ANOVA with list as a between-subjects factor. No significant effects were found ($p > .10$). Familiarity rating was positively skewed and analyzed with the Friedman test. The effect of type on word rating was significant ($\chi^2 = 23.74$, $p < .001$). Despite our attempt to equate the word types on familiarity, using published norms, the participants' own familiarity ratings were slightly higher for categorized words ($M = 6.03$, $SD = 0.99$), lower for emotional words ($M = 5.84$, $SD = 0.92$), and lower still for random

neutral words ($M = 5.71$, $SD = 1.05$). We ran three comparisons of all pairs with a Wilcoxon test, p -corrected for multiple comparisons, and found a significant difference between the categorized neutral and the random neutral words ($p < .001$). We examined the contribution of this effect to the recall performance by selecting a subset of participants ($n = 14$) with no significant difference between their ratings of the three word types ($\chi^2 = 0.16$, $p > .10$). A repeated measures ANOVA examining the recall performance of this subset of participants again showed a significant effect of word type [$F(2,26) = 15.36$, $p < .001$]. Planned contrasts showed that free recall of emotional words was better than that of random neutral words ($p = .001$), with no difference between categorized and random neutral words ($p > .10$).

To examine the memory effect in more detail and to overrule the possibility that the differences in rated familiarity led to the memory effect, we conducted a multiple linear regression analysis of the free recall scores for each word with the predictors rated familiarity, im-

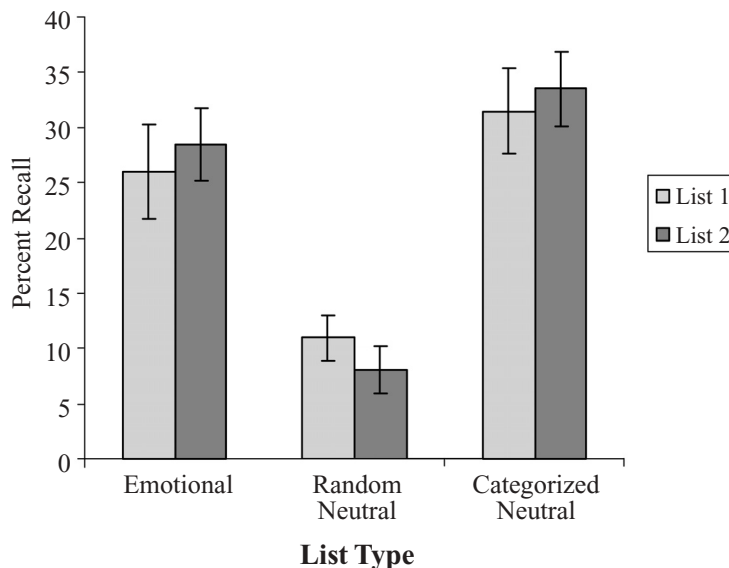


Figure 4. Free recall performance (in percentages) as a function of list type in Experiment 3.

agery, arousal, and relatedness. Rated familiarity was entered first, to partial out any effects it might have had on memory. Entering familiarity rating explained only 5% of the variance ($p > .10$). The overall model explained .57 of the variance [$F(7,46) = 8.86, p < .001$]. Relatedness was a significant predictor of free recall ($\beta = .58; t = 4.22, p < .001$), as well as imagery ($\beta = .35; t = 3.10, p < .005$). No other predictor was significant ($p > .10$).

Discussion

Experiment 3, like Experiments 1 and 2, showed an advantage for emotional items relative to random neutral items. Importantly, this advantage was eliminated when emotional items were compared with categorized neutral items. As in Experiment 2B, which used the same encoding task, here, too, categorized neutral items were recalled better than emotional items. As in Experiments 2A and 2B, relatedness was a significant predictor of recall, whereas emotional arousal was not. As in Experiment 2A, imagery was also a significant predictor of recall.

The additional benefits of enhanced distinctiveness still did not lead to enhanced memory for the emotional words in this experiment. It is possible that the inclusion of categorized neutral items led the participants to rely on relatedness, rather than use distinctiveness to guide their recall, thereby eliminating the advantage of emotional items (see McDaniel et al., 2000, for a similar pattern of findings with bizarre sentences).

GENERAL DISCUSSION

In three experiments, we have shown that the emotionality of words does not make them more memorable than neutral words, when semantic relatedness and imagery are equated for emotional and neutral items. This result suggests that better memory for emotional words than for random neutral words may stem from the fact that emotionality serves as an organizing principle for the items.

The present findings support a suggestion made by previous researchers (Maratos, Allan, & Rugg, 2000; Phelps et al., 1998). In essence, these authors proposed that the improvement of memory performance for emotional material stems from two sources: the arousing nature of the emotional items and their semantic relatedness. However, they did not manipulate semantic relatedness independently of arousal value, so until now, the claim has remained largely untested. In one study, a between-experiments comparison showed that categorized words were recalled better than emotional words, but the emotional and categorized words were not matched on imagery or on semantic relatedness (Doerksen & Shimamura, 2001). Indeed, as was mentioned above, since the lists in that study were drawn from Battig and Montague's (1969) norms, it is likely that the categorized neutral lists were higher on imagery, thus overshadowing any potential effects of emotionality. Another study also seems to

support our conclusions, but the results have never been published (Phelps et al., 1997). Our findings provide direct support for the suggestion that semantic relatedness plays a role in the improvement of memory for emotional material.

Indirect support for the crucial role of semantic relatedness in the enhancement of memory for emotional words can also be found in recognition studies in which memory for emotional and neutral words was compared. Those studies that reported recognition scores corrected for false alarms (Danion, Kauffmann-Müller, Grangé, Zimmermann, & Greth, 1995; Dewhurst & Parry, 2000, Experiment 1; Doerksen & Shimamura, 2001; Maratos et al., 2000; Windmann & Kutas, 2001) showed no enhancement of memory for emotional words, in accordance with the notion that the recognition task provides less opportunity to employ mnemonic organization strategies that rely on relatedness. In contrast to the null effects in word recognition, picture recognition is higher for emotional items (e.g., Bradley et al., 1992; Hamann et al., 1997; Hamann et al., 1999; Ochsner, 2000). Since pictures are presumably more emotionally evocative than words, this is compatible with the possibility that when semantic relatedness is prevented from playing a role, the material has to be very arousing in order to enhance memory.

Enhancement of recognition for emotional words has been found only in a single study, in which the remember-know paradigm was used (Dewhurst & Parry, 2000, Experiment 1). The memory enhancement was manifested as an increase in remember hits, as well as an increase in the false alarm rate. The authors do not attempt to explain the discrepancy between these findings and those in the other studies mentioned above, which also used mixed lists but did not find enhanced recognition of emotional material. Until this finding is replicated, it is difficult to know how to interpret the results.

It is still an open question as to whether the influence of emotion on memory is direct or mediated by other factors. The suggestion that both emotionality and semantic relatedness contribute to memory leaves open the question of what processes are involved. One possibility is that emotionality exerts some of its influence on our memory by acting as an organizing and binding principle at encoding, at retrieval, or both. This principle could operate differently for different kinds of memories. Within sets of items, as is the case in the present study, emotion acted to increase the association between items, which, in turn, promoted item recall, relative to random neutral items. Emotion may also help bind different features of each particular item, which in turn increases source memory for retrieved items (Doerksen & Shimamura, 2001). When narrative materials are involved, emotionality may create a better "story" that is easier to remember (Phelps et al., 1998). Where remote memories are concerned, it might tie together certain features of our experiences (Wright & Nunn, 2000) or our experiences and self-schema (Conway, Pleydell-Pearce, Whitecross,

& Sharpe, 2002), thus facilitating encoding and retrieval of emotional autobiographical memories (e.g., Westmacott & Moscovitch, 2003). One of the effects of emotion on memory may thus be indirect, via organizational influences, in addition to the more direct influences of heightened arousal. When arousal is less intense, the indirect organizational influence of emotion may play the leading role.

Our findings should not be interpreted to mean that emotionality does not enhance memory directly in other situations or with other stimuli. It is possible that memory for materials with a higher emotional value, such as taboo words or emotional scenes, would be enhanced above and beyond the effects of semantic relatedness. Still, our results suggest that the most parsimonious interpretation is that enhanced memory for nontaboo emotional words is influenced wholly or in large part by the semantic relatedness among the items and by the opportunity to implement organizational strategies that such relatedness provides.

REFERENCES

- ADOLPHS, R., DENBURG, N. L., & TRANEL, D. (2001). The amygdala's role in long-term declarative memory for gist and detail. *Behavioral Neuroscience*, **115**, 983-992.
- BATTIG, W. F., & MONTAGUE, W. E. (1969). Category norms of verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology*, **80**, 1-46.
- BLAKE, T. M., VARNHAGEN, C. K., & PARENT, M. B. (2001). Emotionally arousing pictures increase blood glucose levels and enhance recall. *Neurobiology of Learning & Memory*, **75**, 262-273.
- BRADLEY, M. M., GREENWALD, M. K., PETRY, M. C., & LANG, P. J. (1992). Remembering pictures: Pleasure and arousal in memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 379-390.
- BUCHANAN, T. W., DENBURG, N. L., TRANEL, D., & ADOLPHS, R. (2001). Verbal and nonverbal emotional memory following unilateral amygdala damage. *Learning & Memory*, **8**, 326-335.
- COLOMBEL, F. (2000). The processing of emotionally positive words according to the amount of accessible retrieval cues/Traitement de concepts à connotation émotionnelle positive selon la quantité d'indices de récupération accessibles. *International Journal of Psychology*, **35**, 279-286.
- COLTHEART, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, **33A**, 497-505.
- CONWAY, M. A., PLEYDELL-PEARCE, C. W., WHITECROSS, S., & SHARPE, H. (2002). Brain imaging autobiographical memory. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 229-284). San Diego: Academic Press.
- DANION, J. M., KAUFFMANN-MÜLLER, F., GRANGÉ, D., ZIMMERMANN, M. A., & GRETH, P. (1995). Affective valence of words, explicit and implicit memory in clinical depression. *Journal of Affective Disorders*, **34**, 227-234.
- DEWHURST, S. A., & PARRY, L. A. (2000). Emotionality, distinctiveness, and recollective experience. *European Journal of Cognitive Psychology*, **12**, 541-551.
- DOERKSEN, S., & SHIMAMURA, A. P. (2001). Source memory enhancement for emotional words. *Emotion*, **1**, 5-11.
- GUY, S. C., & CAHILL, L. (1999). The role of overt rehearsal in enhanced conscious memory for emotional events. *Consciousness & Cognition*, **8**, 114-122.
- HAMANN, S. B., CAHILL, L., & SQUIRE, L. R. (1997). Emotional perception and memory in amnesia. *Neuropsychology*, **11**, 104-113.
- HAMANN, S. B., ELY, T. D., GRAFTON, S. T., & KILTS, C. D. (1999). Amygdala activity related to enhanced memory for pleasant and aversive stimuli. *Nature Neuroscience*, **2**, 289-293.
- KENSINGER, E. A., BRIERLEY, B., MEDFORD, N., GROWDON, J. H., & CORKIN, S. (2002). Effects of normal aging and Alzheimer's disease on emotional memory. *Emotion*, **2**, 118-134.
- KINSBOURNE, M., & GEORGE, J. (1974). The mechanism of the word-frequency effect on recognition memory. *Journal of Verbal Learning & Verbal Behavior*, **13**, 63-69.
- KLEINSMITH, L. J., & KAPLAN, S. (1963). Paired-associate learning as a function of arousal and interpolated interval. *Journal of Experimental Psychology*, **65**, 190-193.
- LABAR, K. S., & PHELPS, E. A. (1998). Arousal-mediated memory consolidation: Role of the medial temporal lobe in humans. *Psychological Science*, **9**, 490-493.
- MANDLER, G. (1967). Organization and memory. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 1, pp. 327-372). New York: Academic Press.
- MARATOS, E. J., ALLAN, K., & RUGG, M. D. (2000). Recognition memory for emotionally negative and neutral words: An ERP study. *Neuropsychologia*, **38**, 1452-1465.
- MCDANIEL, M. A., DELOSH, E. L., & MERRITT, P. S. (2000). Order information and retrieval distinctiveness: Recall of common versus bizarre material. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 1045-1056.
- OCHSNER, K. N. (2000). Are affective events richly recollected or simply familiar? The experience and process of recognizing feelings past. *Journal of Experimental Psychology: General*, **129**, 242-261.
- PALOMBA, D., ANGRILLI, A., & MINI, A. (1997). Visual evoked potentials, heart rate responses and memory to emotional pictorial stimuli. *International Journal of Psychophysiology*, **27**, 55-67.
- PHELPS, E. A., LABAR, K. S., ANDERSON, A. K., O'CONNOR, K. J., FULBRIGHT, R. K., & SPENCER, D. D. (1998). Specifying the contributions of the human amygdala to emotional memory: A case study. *Neurocase*, **4**, 527-540.
- PHELPS, E. A., LABAR, K. S., & SPENCER, D. D. (1997). Memory for emotional words following unilateral temporal lobectomy. *Brain & Cognition*, **35**, 85-109.
- PUFF, C. R. (1970). Role of clustering in free recall. *Journal of Experimental Psychology*, **86**, 384-386.
- RUBIN, D. C., & FRIENDLY, M. (1986). Predicting which words get recalled: Measures of free recall, availability, goodness, emotionality, and pronounciability for 925 nouns. *Memory & Cognition*, **14**, 79-94.
- TOGLIA, M. P., & BATTIG, W. F. (1978). *Handbook of semantic word norms*. Hillsdale, NJ: Erlbaum.
- TULVING, E., & PEARLSTONE, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning & Verbal Behavior*, **5**, 381-391.
- WESTMACOTT, R., & MOSCOVITCH, M. (2003). The contribution of autobiographical significance to semantic memory. *Memory & Cognition*, **31**, 761-774.
- WINDMANN, S., & KUTAS, M. (2001). Electrophysiological correlates of emotion-induced recognition bias. *Journal of Cognitive Neuroscience*, **13**, 577-592.
- WRIGHT, D. B., & NUNN, J. A. (2000). Similarities within event clusters in autobiographical memory. *Applied Cognitive Psychology*, **14**, 479-489.

(Manuscript received June 3, 2003;

revision accepted for publication January 8, 2004.)