Cogn Process (2013) 14:73–79
DOI 10.1007/s10339-012-0530-4

SHORT REPORT

Does category labeling lead to forgetting?

Nathaniel Blanco · Todd Gureckis

Received: 8 March 2012 / Accepted: 22 August 2012 / Published online: 12 October 2012

Abstract What effect does labeling an object as a member of a familiar category have on memory for that object? Recent studies suggest that recognition memory can be negatively impacted by categorizing objects during encoding. This paper examines the "representational shift hypothesis" which argues that categorizing an object impairs recognition memory by altering the trace of the encoded memory to be more similar to the category prototype. Previous evidence for this idea comes from experiments in which a basic-level category labeling task was compared to a non-category labeling incidental encoding task, usually a preference judgment (e.g., "Do you like this item?"). In two experiments, we examine alternative tasks that attempt to control for processing demands and the degree to which category information is explicitly recruited at the time of study. Contrary to the predictions of the representational shift hypothesis, we find no evidence that memory is selectively impaired by category labeling. Overall, the pattern of results across both studies appears consistent with well-established variables known to influence memory such as encoding specificity and distinctiveness effects.

Keywords Categorization · Labeling · Memory · Schema encoding

N. Blanco (✉)
Department of Psychology, University of Texas at Austin,
108 Dean Keaton Stop A8000, Austin, TX 78712, USA
e-mail: nathanblanco@gmail.com

T. Gureckis
Department of Psychology, New York University,
6 Washington Place, New York, NY 10003, USA

Introduction

Since the seminal studies of Bartlett (1932), the idea that memory and perception can be deeply influenced by prior conceptual knowledge has been well appreciated. One way in which everyday knowledge makes contact with experience is through categorization. Categorization is a critical cognitive ability which allows us to recognize novel objects in our environment as coming from distinct classes. However, the full relationship between categorization and memory remains poorly understood. For example, what effect does categorizing an object with respect to an established knowledge structure have on subsequent memory for the object?

Interestingly, a number of recent studies have shown that categorization can sometimes have a negative impact on subsequent recognition memory. For example, Sloutsky and Fisher (2004) found that 5-year-olds who perform an induction task on a set of stimuli exhibit better memory for those objects than college-aged adults. The hypothesis advanced by Sloutsky and Fisher was that adults perform induction based on prior knowledge of real-world categories leading to more general, category-level representations of the presented objects, while the younger individuals (who lack such general world knowledge) use a more perceptual, similarity-based strategy.

One of the most provocative demonstrations of the effect that categorization has on subsequent memory is a recent set of findings reported by Lupyan (2008). In this study, participants viewed photographs of everyday household objects while performing one of two tasks (Experiment 4). On some trials, an object was presented on the screen briefly, and the participant was asked to label the object according to its basic-level category (e.g., by determining whether it was a "chair" or a "lamp").

On other trials, a preference judgment was elicited ("Do you like this item? Yes or No?"). Following an initial study phase where these two types of trials were randomly intermingled, participants were given a recognition memory test. The study found that recognition performance was lower for the items that had been labeled relative to those for which a preference judgment had been elicited.

Lupyan (2008) explained this result in terms of a "representational shift." According to this account, verbally labeling an object according to its basic-level category activates the high-level category representation (i.e., prototype) which then exerts a top-down influence on the encoding process. This additional activation has the effect of shifting the representation of the study item toward the category prototype. When the studied object is seen again, there is a greater mismatch between the perceptual experience of the object and the representation stored in memory, making participants less likely to recognize it as seen previously. The idea that top-down activation from verbal processes might distort memory processing shares some similarity to accounts of the verbal-overshadowing effect whereby verbally describing an experience degrades memory for the specific details of the event (Schooler and Engstler-Schooler 1990).

Does labeling cause forgetting? If so, why?

Intriguing as such demonstrations are, the interpretation of the results is, as yet, unclear. While the experiments clearly show that category labeling results in worse memory (relative to a non-labeling preference judgment task), there is an evidence to suggest that preference judgments might simply result in better memory relative to other tasks (Richler et al. 2011). For example, participants may have found the categorization task easy since all the objects involved were highly discriminable real-world objects and thus may not have spent much time considering the individual features of each object. In contrast, it seems likely that preference judgments might result in deeper consideration of the idiosyncratic features of an object (e.g., "Would this chair match my desk? Would I like to sit on it?"). Similarly, the preference judgment involves emotional processing in a way that categorization does not and which may have facilitated subsequent memory (Kensinger and Corkin 2003). Also note that numerous studies have demonstrated a memory advantage for encoding items in relation to the self (see Symons and Johnson 1997 for a review). Consistent with this view, Lupyan (2008) found that response times were typically longer for preference judgments than for labeling trials, indicating that participants may have been processing the items more in more detail. Finally, the pattern of results may be an effect of encoding specificity; because preference judgments involve more detailed processing, those details are more effective as retrieval cues during recognition (Tulving and Thompson 1973).

Understanding the nature of this effect is critical. If the pattern of results in the previous work is best explained as a more detailed or deeper encoding of items when given preference ratings, it would undermine the hypothesized role that linguistic labeling has on recognition memory. Instead, the results may be explained in terms of more traditional variables known to influence memory.

In the present paper, we extend the design of Lupyan (2008) to disentangle the role of a top-down/representational shift effect relative to other, well-established memory effects. In Experiment 1, we introduce a second comparison task closely matched to the encoding demands of category labeling while avoiding the interpretation problems associated with preference judgments.

To foreshadow, the results suggest that compared to preference judgments, worse memory also accompanies tasks such as a simple orientation discrimination, which are unlikely to tap the same kind of high-level prototypical features that basic-level category labels are expected to activate. In Experiment 2, we further examine a prediction of the representational shift hypothesis, namely that factors that more strongly activate category knowledge should more strongly impair memory. In this experiment (a conceptual replication of Lupyan's Experiment 5), we fail to find a robust difference between categorization-relevant and categorization-irrelevant study conditions. However, subsequent analysis suggests that memory performance is strongly influenced by the distinctiveness of study items during study (items given more extreme ratings of typicality or preference are better remembered). Together, these results cast doubt on the idea that category labeling has a specific effect on recognition memory above and beyond factors that are known to influence encoding and retrieval.

Experiment 1

Experiment 1 is a conceptual replication and extension of Lupyan (2008), Experiment 4 (described above). However, we introduce a third study task which requires participants to judge the orientation of the presented object (facing left or right). The orientation task has a number of properties that make it a desirable control for both the preference and category labeling tasks. First, unlike the preference judgment task, the orientation discrimination requires little processing of specific idiosyncratic features of the items. In addition, while orientation (left/right) may be seen as a type of categorization judgment, it is unlikely to activate the same kind of prototypical features as labeling an object

according to its basic-level category. Finally, the orientation task avoids the issues with self-schema encoding (Symons and Johnson 1997) or emotional processing which accompany preference judgments (Kensinger and Corkin 2003). Therefore, this task provides a good baseline to determine which of these two potential effects contribute to the pattern of results found in the original study. If a representational shift were to occur for labeled items, we expect them to be recognized less often than items in the orientation condition. If the preference task results in particularly deep encoding, then preference items should be remembered better than both labeled and orientation items. Of course, it is also possible for both, or neither, effects to occur.

Methods

Participants and apparatus

Forty students at New York University participated in partial fulfillment of a class requirement. The mean age of the population from which participants were recruited was 19 years old, with approximately 66 % females. The experiment was administered on standard Macintosh computers over a single half-hour session.

Stimuli

The stimuli were photographs of individual everyday objects isolated against a plain white background. Images of chairs and lamps were used for the main experiment, and images of clocks and bookshelves were used for practice trials. The photographs were obtained from online catalogs, the majority from the IKEA online catalog (www.ikea.com). Most of the stimuli were from the set used by Lupyan (2008), obtained through personal correspondence. However, stimuli that did not have a clear orientation were replaced with related alternatives. The stimuli were divided into two main sets. Each object in a given set was matched to an object in the other set that was highly similar in appearance and acted as its critical lure (see Fig. 1 for examples). Each set consisted of 20 chairs and 20 lamps. One half of the sets of chairs and lamps were oriented facing to the left, and the other half were facing the right.

Procedure

The experiment consisted of two sub-experiments, referred to as Experiment 1A and 1B. Experiment 1A was a direct replication of Lupyan's (2008), Experiment 4. Experiment 1B was identical to Experiment 1A, except that the labeling task was replaced by an orientation judgment task. The experiments consisted of practice trials, a study phase, and a subsequent recognition memory test. The study phase included only the labeling and preference tasks in Experiment 1A, and the orientation and preference tasks in Experiment 1B.

In the category labeling task, subjects indicated whether the object was a chair or lamp. In the preference task, subjects indicated whether they liked the object or not. In the orientation task, subjects indicated whether the object was facing to the left or to the right. Prior to the study phase, participants were given 48 practice trials using images of clocks and bookshelves to acclimate them to the tasks.

On each trial of the study phase, a photograph of an object was presented in the center of the screen for 300 ms. Then a mask was displayed for 300 ms. The mask disappeared and immediately following, a prompt was displayed indicating which of the two tasks to perform. From the onset of the prompt, subjects were given a 2-s window in which to respond by pressing an appropriate key on the keyboard. After an inter-trial interval of 1.5 s, subjects hit the space bar whenever they were ready to begin the next trial. The two tasks were randomly intermingled.

One of the two stimulus sets was randomly selected as the study set for each subject. Each stimulus from that set was presented twice over the course of the study phase (each time paired with the same task). Across participants, which task was performed on which specific stimuli was randomly counterbalanced.

In the test phase, participants were presented with a stimulus and instructed to indicate whether they saw the image during the study phase or not. The image remained on the screen until the subject made a response by pressing a key on the keyboard. All of the stimuli from both sets were presented (in random order) once during the test: the 40 old items viewed during the study phase and the 40 novel items from the other set (the critical lures).

Results

Performance was high on both the labeling task and orientation tasks, 99 % correct and 97 % correct, respectively. In the preference task, items were liked slightly more often than disliked, with subjects giving a "like" response to an average of 55 % of the items.

Overall memory performance (independent of encoding condition) did not differ between Experiment 1A and 1B on measures of d', hit rate, or false alarm rate ($t(38) < 1$ for all three comparisons). Collapsed across experiments, overall d' was 0.97, hit rate was 0.67, and false alarm rate (endorsing the critical lures) was 0.32.

Our key dependent measure was memory performance (indicated by d') during the test phase as a function of encoding task. Table 1 summarizes our key findings.

Fig. 1 Example stimuli for Experiment 1 and 2. Targets were seen in the study phase. Both targets and lures were presented (independently) during the recognition test. Which items were targets and which lures was counterbalanced between participants

Table 1 Summary of the main dependent measures for Experiment 1 and 2 compared with the results from Lupyan (2008), Experiment 4 and 5

Experiment/condition	Hits	False Alarms	d'	RT (ms)	Cohen's d
Experiment 1A	*		*	*	
Preference	.72 (.17)	.32 (.12)	1.14	697 (142)	.54
Category labeling	.63 (.15)	.34 (.17)	.85	568 (72)	
Experiment 1B	*	*	*	*	
Preference	.74 (.15)	.34 (.13)	1.16	868 (172)	.52
Orientation	.60 (.15)	.29 (.15)	.88	634 (121)	
Lupyan (2008) Experiment 4	*				
Preference	.71 (.09)	.32 (.14)	1.11		.76
Category labeling	.62 (.14)	.40 (.19)	.64		
Experiment 2					
Preference (1–5)	.76 (.13)	.30 (.13)	1.31	1092 (232)	.05
Typicality (1–5)	.75 (.14)	.29 (.12)	1.34	1062 (233)	
Lupyan (2008) Experiment 5	*		*	*	
Preference (y/n)	.83 (.10)	.41 (.16)	1.29	753 (121)	.59
Typicality	.75 (.13)	.40 (.16)	1.00	978 (119)	

The mean hit rate and false alarm rate are shown (standard deviations in parentheses), as well as d', RT, and Cohen's effect size measure (d) calculated on the difference in d' between conditions. A star in a column indicates that the two conditions differed significantly ($p < .05$) on that measure in the experiment. Lupyan (2008) did not report RT for Experiment 4

Experiment 1A replicated the results found by Lupyan (2008). Performance (d') was lower for labeled items, $t(19) = 2.27$, $p = 0.04$. Hit rates also differed systematically as a function of encoding task, $t(19) = 2.29$, $p = 0.03$. False alarm rates did not differ by encoding task, $t(19) < 1$.

Results for Experiment 1B were qualitatively similar to Experiment 1A. Again, performance (d') was higher for the items studied under the preference judgment task, $t(19) = 2.50$, $p = 0.02$. Hit rates were also higher for the preference items, $t(19) = 3.86$, $p < 0.005$. Though, unlike Experiment 1A, false alarm rates were higher for the preference items than for the orientation, $t(19) = 2.42$, $p = 0.03$.

Importantly, performance in the labeling condition in Experiment 1A and the orientation condition in Experiment

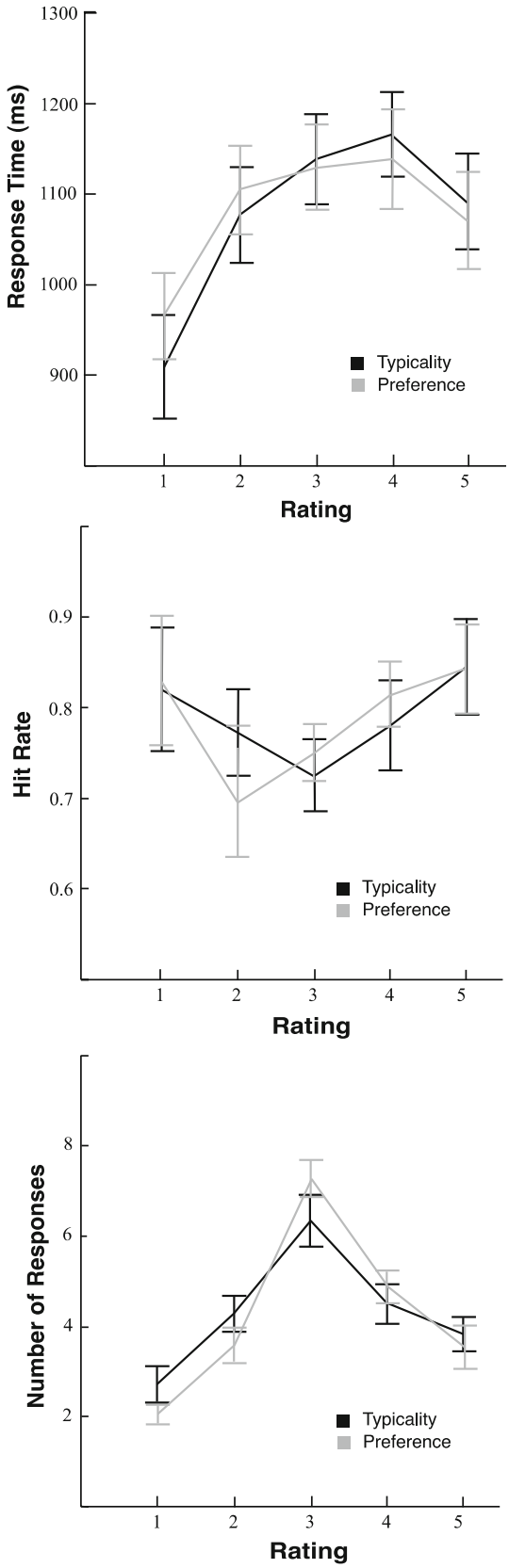

Fig. 2 *Top*: RT during the study phase of Experiment 2 as a function of rating given. *Middle*: Hit rate in the test phase as a function of rating given during the study phase. *Bottom*: The average number of times each rating was given. For both tasks items were given extreme values less often, and these items have the shortest response times and the highest hit rates. *Error bars* reflect SE

1B did not differ significantly in *d'*, $t(38) < 1$; hit rate, $t(38) < 1$; or false alarms, $t(38) < 1$. Memory performance for the preference items also did not differ between Experiments 1A and 1B, $t(38) < 1$ for all measures.

In each experiment, the pattern of RT in the study phase mirrors the pattern of hit rates in the test phase, suggesting that longer RT leads to higher hit rates. RT during the preference task was significantly longer than during the category labeling task (Experiment 1A), $t(19) = 4.87$, $p < 0.005$, and orientation (Experiment 1B) task, $t(19) = 6.19$, $p < 0.005$.

Discussion

We found that memory performance following an orientation judgment task was reduced relative to the preference judgment condition, but equivalent to the category labeling condition. Presumably, the orientation judgment task did not require activation of the category prototype at the time of study (at least no more so than for preference judgments). The pattern of memory deficits suggests that the labeling and orientation tasks had nearly identical effects on subsequent memory, thereby raising questions about the representational shift hypothesis. Overall, preference judgments appear to invoke deeper processing of the target object, and as a result, memory is improved for these items. In contrast, the orientation and labeling tasks were both simple judgments that could be made without deeply processing the specific perceptual details of the object.

Experiment 2

While encoding differences between category labeling and preference judgments seem to account for the results of our experiment and those of Lupyan (2008), Experiment 1–4 (which consistently found that RTs were longer for preference judgments than category labeling), another experiment from Lupyan (2008) provides a different kind of support for the representational shift hypothesis. In particular, in Lupyan's (2008) experiment, Experiment 5, memory performance was compared in two conditions. In the first task, participants were asked to give typicality ratings (on a scale from 1 to 5) to items, and in the second task participants gave binary (like/dislike) preference judgments.

Interestingly, Lupyan found that response times for typicality ratings had a nonlinear pattern such that ratings

of both highly typical and highly atypical items were made quickly while items of intermediate typicality were judged more slowly. Hit rates showed an inverse pattern: Hit rates were highest for the most and least typical items (i.e., those with the shortest RT). Lupyan explained the pattern of results as being characteristic of a representational shift. According to this idea, during typicality judgments, more time spent processing the item in relation to the category might cause a stronger top-down influence on encoding. More category-relevant processing leads to poorer subsequent memory. As a result, items with the longest RTs (i.e., those associated with intermediate ratings) should have the lowest hit rates. Since this pattern is hypothesized to be a result of a representational shift, it should be unique to category-related encoding tasks, and therefore, you would not expect it from other tasks, such as preference judgments.

In Experiment 2, we test this implication directly in order to investigate the nature of this unusual effect. We use the same basic tasks as Lupyan's Experiment 5, typicality and preference judgments, but we equate the scale used for the two tasks.

Methods

Participants and apparatus

Twenty-nine students at New York University participated in partial fulfillment of a class requirement. The general population characteristics were the same as in Experiment 1. The experiment was administered on standard Macintosh computers over a single half-hour session. The stimuli were the same as those used in Experiment 1.

Procedure

As in Experiment 1, the experiment consisted of practice trials, a study phase, and a recognition test. The recognition test was identical to that of Experiment 1. The procedure for the study phase and practice trials was the same as Experiment 1 with the following exceptions. In the typicality task, subjects were asked to indicate on a scale from 1 to 5 how typical the object is for its category (e.g., "How typical is this lamp?" 1 = very typical, 5 = very atypical), and in the preference task, subjects indicated how much they like the object on a scale from 1 to 5 (1 = really like, 5 = really dislike).

Results

Across subjects overall *d'* independent of encoding condition was 1.31. Overall hit rate was 0.76 and the false alarm rate 0.30. Memory performance (*d'*) did not differ between items studied under the two different tasks, $t(28) < 1$ (see Table 1). Hit rate and false alarms also did not differ by task, $t(28) < 1$ for both comparisons.

RT for typicality judgments ($M = 1062$ ms, $SD = 233$ ms) was slightly shorter than RT for preference judgments ($M = 1092$ ms, $SD = 232$ ms), $t(28) = 2.11$, $p = 0.04$. As shown in Fig. 2 (top), RTs for both typicality and preference judgments followed an approximately inverted U-shaped curve based on the rating given, with intermediate ratings producing longer RTs.

For each condition we analyzed hit rate as a function of the ratings given during the study phase. For each subject, the average typicality or preference rating was computed for each item (since each item was seen twice), and the items were placed into bins based on that rating. Figure 2 shows the average hit rate for each of the bins for both conditions. Consistent with Lupyan's (2008) findings, the hit rates for items studied under the typicality task form a roughly U-shaped pattern. Critically though, hit rates for the preference items follow the same pattern.

One explanation of the U-shaped memory effect in both the preference and typicality rating conditions is that fewer items were given extreme scores of either 1 or 5. Figure 2 (bottom) shows the average number of items given each rating score for both types of encoding tasks. Critically, fewer items were given extreme ratings (i.e., a score of 1 or 5). As a result, it is plausible that these items were better differentiated in memory relative to the larger number of items that were given intermediate ratings. To evaluate this hypothesis, we performed an ANCOVA with the number of study items given each rating as a covariate and the actual rating the item was given (1-5) as a factor (collapsed across condition). We found a significant effect of the covariate ($F(1,251) = 4.53$, $p = .03$), but no effect of the rating itself ($F(4,251) = 2.04$, $p = .09$) nor an interaction between these variables ($F(4,251) = 0.68$, $p = .6$). The beta weights estimated for the covariate were all negative (reflecting the negative relationship between the number of items within a bin and the hit rate).

Discussion

According to the representational shift hypothesis, the deeper category-related processing required by typicality judgments should cause a larger decrement in memory performance compared to the preference judgment task. In our experiment which equated the response scale used for the two tasks, we found that memory performance did not differ as a function of study task. Alone this would appear to be a null effect. However, we replicated the U-shaped pattern of hit rates as a function of typicality rating found in Lupyan (2008), Experiment 4, as well as the inverted U-shape for RT. Importantly, we found identical patterns

for items studied under the preference rating task, suggesting that this pattern is not unique to categorical processing. Closer analysis of our data showed that there were consistently fewer items given extreme ratings in both tasks. A plausible explanation of this effect is that the items at the end points of the scale are given enhanced encoding by virtue of being relatively unique. This was supported by our ANCOVA which found that the number of items given each rating was a better predictor of hit rates than the rating itself. Ultimately, the results of both our experiment and Lupyan (2008) might be best explained in terms of a distinctiveness effect (von Restorff 1933; Sakamoto and Love 2006; Nairne 2002). Items were rarely given extreme values on the rating scale, making them more distinct in memory.

General discussion

In this paper, we examined recent findings that indicated that labeling something as a member of a familiar category might result in a decrement in memory for that item (Lupyan 2008). The goal of this paper was simply to determine whether category labeling impairs memory separately from these other factors. In Experiment 1, we showed that we can produce similarly reduced memory performance for other tasks that do not obviously overlap with basic-level category labeling. Our data suggest that a critical feature of the original effect may not be that labeled items are remembered poorly as much as preference judgments result in superior memory relative to a variety of incidental encoding tasks. We interpreted this in terms of encoding specificity effect (Tulving and Thompson 1973), whereby detailed processing of the items created associated retrieval cues which aided subsequent memory.

Experiment 2 examined a secondary finding in Lupyan (2008), an interesting pattern of RT and subsequent hit rate for items given typicality ratings. To investigate to what extent the specific pattern of results found is unique to category-relevant tasks (and therefore indicative of representational shift), we replicated the basic features of this design while equating the scale for the typicality and preference rating tasks. In our experiment we found no difference in memory performance for items studied under the two rating tasks. We also showed that the detailed pattern of results is partially explained by the number of items given each rating. Items given extreme ratings of either preference or typicality were relatively infrequent and thus may have stood out better in memory. This hypothesis is, in fact, a straightforward extension of the encoding specificity argument made for the results of Experiment 1. The distinctiveness of the items given extreme ratings may have activated idiosyncratic features (e.g., "that chair is like my favorite chair," "that chair looks very uncomfortable") which served as additional retrieval cues during recognition.

Memory and categorization are fundamentally intertwined processes, and there is no doubt that semantic memory can strongly modulate memory encoding and retrieval processes. Studies such as Sloutsky and Fisher (2004) highlight the powerful influence that category-related processing can have on memory. However, just as fundamental is the fact that memory is influenced by a variety of factors including the processing demands of the encoding task and the context in which items are studied (Tulving and Thompson 1973). We found no evidence that simply labeling an object by its basic-level category uniquely affects subsequent recognition relative to variables traditionally known to influence memory.

Acknowledgments The preliminary results for this study were presented as part of a final class project in an undergraduate Lab in Human Cognition course at NYU. We especially thank Kate Ray and Frank Lei. We also thank Eric Dewitt, members of the Davachi lab, members of the Daw lab, and the Concepts and Categories (ConCats) group at NYU for helpful discussions in the development of this project.

References

Bartlett F (1932) Remembering: an experimental and social study. Cambridge University Press, Cambridge

Kensinger E, Corkin S (2003) Memory enhancement for emotional words: are emotional words more vividly remembered than neutral words? Mem Cogn 31(8):1169–1180

Lupyan G (2008) From chair to "chair:" a representational shift account of object labeling effects on memory. J Exp Psychol Gen 137(2):348–369

Nairne J (2002) The myth of the encoding-retrieval match. Memory 10:389–395

Richler J, Gauthier I, Palmeri T (2011) Automaticity of basic-level categorization accounts for labeling effects in visual recognition memory. J Exp Psychol Learn Mem Cogn 37(6):1579–1587

Sakamoto Y, Love B (2006) Enhanced oddball memory through differentiation, not isolation. Psychonomic Bull Rev 13:474–479

Schooler J, Engstler-Schooler T (1990) Verbal overshadowing of visual memories: some things are better left unsaid. Cogn Psychol 22:36–71

Sloutsky V, Fisher A (2004) When development and learning decrease memory. Psychol Sci 15(8):553–558

Symons C, Johnson B (1997) The self-reference effect in memory: a meta-analysis. Psychol Bull 121:371–394

Tulving E, Thomson D (1973) Encoding specificity and retrieval processes in episodic memory. Psychol Rev 80(5):352–373

von Restorff H (1933) Analyse von vorgangen in spurenfeld. i. uber die wirkung von bereichs- bildungen im spurenfeld [analysis of processes in the memory trace on the effect of group formations on the memory trace]. Psychologische Forschung 18:299–342