Does category labeling lead to forgetting?

Nathaniel Blanco (nathanblanco@gmail.com)
Todd M. Gureckis (todd.gureckis@nyu.edu)
New York University, Department of Psychology
6 Washington Place
New York, NY 10003 USA

Abstract

In this paper, we evaluate the "representational shift" hypothesis (Lupyan, 2008) which argues that the act of explicitly labeling an object as a member of a familiar semantic category alters the trace of the encoded memory in the direction of the category prototype. The typical procedure for such experiments has been to compare category labeling to a non-categorization encoding task such as a preference judgement. In a series of experiments, we examine alternative comparison tasks that attempt to control the depth of encoding and the degree to which category information is explicitly recruited at the time of study. The results appear most consistent with a depth of processing (Craik & Lockhart, 1972) (Exp. 1) or distinctiveness (Exp. 2) explanation for the pattern of memory effects found in previous studies.
Keywords: categorization, labeling, memory, depth of processing, schema-encoding

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. Other studies have shown that the very act of learning a category can fundamentally distort perceptual representations (Harnad, 1987; Goldstone & Hendrickson, 2010).

One of the most provocative demonstration 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 objects (such as chairs, lamps and tables) while performing one of two tasks (Exp. 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 (for example, by determining if the item was a "chair" or a "lamp"). On other trials, a preference judgement was elicited ("Do you like this item? Yes or No?"). Following an initial phase where these two types of trials were randomly intermingled, participants were given a recognition memory test. During the test, participants were shown the original items along with a set of highly similar lures and for each item were asked to report if the item was presented during the initial study phase. The study found that recognition performance was lower for the items that had been labeled relative to those for which a preference judgement had been elicited.

Lupyan (2008) explained this results in terms of a "representational shift" hypothesis (also referred to as the label-feedback hypothesis). According to this account, verbally labeling an object according to its basic-level category activates the high-level category representation (i.e., prototype) which in turn 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 during the recognition test, there is a greater mismatch between the perceptual experience of the object and the representation stored in memory, make subjects 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 & 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, relative to a non-labeling task (preference judgements), category labeling results in worse memory, it is also possible that preference judgements simply result in better memory relative to other tasks. 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 judgements might result in deeper consideration of the idiosyncratic features of an objects (e.g., "Would this chair match my desk? Would I like to sit on it?"), something Lupyan acknowledged in his original report. 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 deeply. In addition, numerous studies have demonstrated a memory advantage for encoding items in relation to the self (see Symons & Johnson, 1997 for a review).

The goal of the present study is to further explore the impact that categorization or labeling has on memory processes. In particular, we extend the design of Lupyan (2008) in a way that would allow us to disentangle the role that depth-of-processing might have played relative to a top-down conceptual/representational shift effect. In Experiment 1, we replicate the results of the original study while introducing a second comparison task which is closely matched to the demands of category labeling while avoiding the interpretation problems associated with preference judgements. To foreshadow, the results suggest that 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.

Experiment 1

Experiment 1 is a conceptual replication and extension of Luypan (2008), Experiment 4 (described above). We introduced a third study task which required participants to judge the orientation of the presented object (facing left or facing right). The orientation task has a number of properties that make it a desirable control for both the preference and category labeling task. First, unlike the preference judgement 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 judgement, it is unlikely to activate the same kind of prototypical features as labeling an object according to its basic-level category. Orientations judgments should neither result in reduced memory performance caused by a representational shift nor improved performance caused by requiring attention detail. Therefore, it provides a good baseline to determine which of these two potential effects contribute to the pattern of results in the original study. If a representational shift occurs for labeled items we expect lower memory performance for these items compared to those in the orientation condition. On the other hand, if the preference task causes particularly good memory, then preference items should be remembered better than both labeled items and orientation items.

Figure 1: Example stimuli for Experiments 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.

Methods

Participants and Apparatus 40 students at New York University participated in partial fulfillment of a class requirement. 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. The photographs used for the main experiment were pictures of chairs and lamps. In addition, six images of clocks and six images of bookshelves were used during practice trials. The photographs were obtained from online catalogs, the majority from the IKEA online catalog (www.ikea.com). The majority of the stimuli were from the set used by Lupyan (2008), obtained through personal correspondence. However, stimuli that did not have 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 Figure 1 for examples). Each set consisted of 20 chairs and 20 lamps. One half (i.e., 10) of the chairs/lamps were oriented facing to the left, and the other half were facing the right.

Table 1: Summary of the main dependent measures for Experiment 1 and 2 compared with the results from Lupyan (2008), Exp. 4 and 5. The mean hit rate and false alarm rate is shown (standard deviations in parentheses), as well as d' and Cohen's effect size measure (d) calculated on the difference in d' between conditions.

Experiment/Condition	**Hits**	**False-Alarms**	d'	**Cohen's** d
Experiment 1A	*		*	
Preference	.72 (.17)	.32 (.12)	1.14	.54
Category Labeling	.63 (.15)	.34 (.17)	.85	
Experiment 1B	*	*	*	
Preference	.74 (.15)	.34 (.13)	1.16	.52
Orientation	.60 (.15)	.29 (.15)	.88	
Lupyan (2008) Exp. 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	.05
Typicality (1-5)	.75 (.14)	.29 (.12)	1.34	
Lupyan (2008) Exp. 5	*		*	
Preference (y/n)	.83 (.10)	.41 (.16)	1.29	.59
Typicality (1-5)	.75 (.13)	.40 (.16)	1.00	

* $p < .05$

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), Exp. 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 the labeling and preference tasks in Exp. 1A, and the orientation and preference tasks in Exp. 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 (the instructions made clear how to interpret this task and stimuli were selected so that the orientation was always obvious).

Before the study phase, participants were given 48 practice trials to acclimate them to the timing of the tasks. The practice trials were identical to the trials of the study phase, except images of clocks and bookshelves were shown instead of the stimuli used in the main experiment.

On each trial of the 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 second window in which to respond. The prompt remained on the screen until the subject made a response by pressing an appropriate key on the keyboard or until the two seconds expired, whichever came first.

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 once during the test phase: the 40 old items viewed during the study phase and the 40 novel items from the other set (the critical lures). The test stimuli were presented in random order.

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 Exp. 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. Exp. 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 Exp. 1B were qualitatively similar to Exp. 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 Exp. 1A, false alarm rates were marginally higher for the preference items than orientation, $t(19) = 2.42$, $p = 0.03$.

Importantly, performance in the labeling condition in Exp. 1A and the orientation condition in 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.

Response times (RT) during the study phase were analyzed to assess their effect on the memory results. 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 lead to higher hit rates. RT during the preference task was significantly longer than during the category labeling (Exp. 1A), $t(19) = 4.87$, $p < 0.005$, and orientation (Exp. 1B) task, $t(19) = 6.19$, $p < 0.005$.

Discussion

We found that memory performance following an orientation judgement task was reduced relative to the preference judgement condition, but equivalent to the category labeling condition. Presumably, the orientation judgement task did not require activation of the category prototype at the time of study (at least relative to the category labeling condition). The pattern of memory deficits suggest that the labeling and orientation tasks had nearly identical effects on subsequent memory, thereby undermining the representational shift hypothesis. The correlation between study response time and subsequent memory is consistent with a levels-of-processing hypothesis (Craik & Lockhart, 1972). Overall, preference judgements 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 judgements that could be made without deeply processing the specific perceptual details of the object.

Experiment 2

While depth-of-processing differences between category labeling and preference judgments would appear to account for the results of our experiment and the results of Lupyan (2008) Exp. 1-4 (which also consistently found that response times were longer following preference judgements than in category labeling condition), Lupyan presented at least one finding which would appear to rule out this straight-forward interpretation. In particular, in a followup experiment (Exp. 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 yes/no preference judgements. The study found that the typicality judgment elicited both longer RTs and lower hit rates than the binary preference judgment. While longer RT does not always mean deeper encoding, there is no obvious depth-of-processing explanation for this result.

Interestingly, Lupyan also found that response times for typicality ratings had a non-linear 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). This result also seems to rule out a simple depth-of-processing explanation. Lupyan explained the pattern of results as being characteristic of a representational shift. According to this idea, in the typicality judgement task, more time spent processing the item in relation to the category might cause a stronger top-down influence on encoding. As a result, typicality items with the longest RTs (i.e., those associated with intermediate ratings) should have the lowest hit rates. An import implication of this explanation is that this pattern of results should be unique to category-related encoding tasks such as typicality judgments, but *not* preference judgments.

In Experiment 2, we test this implication directly. We use the same basic tasks as Lupyan's Exp. 5, typicality judgments and preference judgments, but we simply equated the scale used for the two tasks.

Methods

Participants and Apparatus 29 students at New York University participated in partial fulfillment of a class requirement. The experiment was administered on standard Macintosh computers over a single half-hour session. The stimuli were the same as those used in Exp. 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 Exp. 1. The procedure for the study phase and practice trials was the same as Exp. 1 with the following exceptions. In

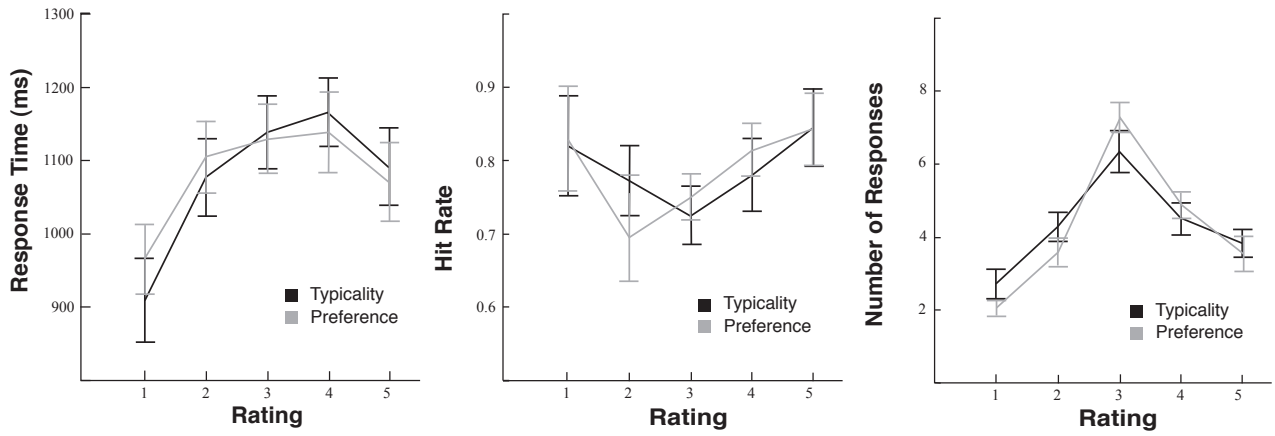

Figure 2: *Left*: RT during the study phase of Exp. 2 as a function of rating given. *Middle*: Hit rate in the test phase as a function of rating given during the study phase. *Right*: The average number of times each rating was given. For both tasks items were given extreme values less often, and these items have the shortest response times and the highest hit rates.

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?" where 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 (where 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. Again, the key dependent measure was memory performance (indicated by d') during the test phase as a function of the encoding task (Table 1). Memory performance (d') did not differ between items studied under the two different tasks, $t(28) < 1$. Additionally, neither hit rate nor false alarms differed as a function of encoding task, $t(28) < 1$ for both comparisons.

RT for typicality judgments (M=1062 ms, $SD = 233$ ms) was slightly shorter than RT for preference judgements ($M = 1092$ ms, $SD = 232$ ms), $t(28) = 2.11$, $p = 0.04$. As shown in Figure 2 (left) RTs for both typicality and preference judgements followed an approximately inverted U-shaped curve based 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 (each item was seen twice), and the items were placed into bins based on that average 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 (right) show 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 number of study items given each rating as a covariate and the actual rating (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 ($F(4, 251) = 0.68$, $p = .6$). In addition, the beta weights estimated for the covariate were all negative (reflecting the negative relationship between 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 judgement task. In our experiment which equated the response scale used during both tasks, we found that memory performance did not differ as a function of typicality or preference judgments. Alone this would appear to be a null effect. However, we replicated the U-shaped pattern of hit rates as a function typicality rating that was found in Lupyan (2008, Exp. 4), as well as the inverted U-shape for RT. In particular, items given extreme typicality ratings have higher hit rates (and lower RT) than those in the center of the scale. Importantly, we also found an identical pattern for items studied under the 1-5 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. One 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 analysis which found that the number of items given each rating was a better predictor of hit rates than was 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 & Love, 2006). Items were rarely given extreme values on the rating scale, making them more distinct in memory.

General Discussion

In this paper, we examined a recent study which found that labeling something as a member of a familiar category can result in a decrement in memory for that item (Lupyan, 2008). 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 (e.g., judging the orientation of an familiar object). Our data suggest that a critical feature of the original results may not be that labeled items are forgotten as much as making preference judgments for objects results in superior memory relative to a variety of other incidental encoding tasks (Symons & Johnson, 1997).

Experiment 2 examined a secondary finding in Lupyan (2008) which found that typicality ratings could also results in lower memory compared to a preference judgement task. This finding was central to the original study because it showed that memory could be worse for items which were associated with longer study RTs (appearing to undermine a simple depth-of-processing account). We replicated the basic features of this design but showed that the detailed pattern of hit rates is best explained in terms of the number of items appear at each rating. Items given extreme ratings of both preference and typicality were relatively infrequent, and thus may have stood out better in memory.

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 recognition that memory is influenced by a variety of factors including the context in which items are studied (Tulving & Thomson, 1973) and the degree or depth of encoding (Craik & Lockhart, 1972). Recently, a number of authors have argued for specific changes in memory or perception based on the top-down influence of linguistic or verbal processing (Schooler & Engstler-Schooler, 1990; Winawer, Witthoft, Frank, Wu, & Wade, 2007; Lupyan, 2008). While we cannot rule out the possibility that category induced distortions in memory (such as those proposed by the representational shift hypothesis) might occur specifically during linguistic or semantic processing, we found no unique evidence to support this hypothesis outside of more traditional variables known to influence memory. At a behavioral level, establishing that one particular type of encoding task retards memory in a distinct manner is challenging. In our view, the best way to make progress on such issues is by leveraging multiple sources of data including insight from cognitive neuroscience about the role of semantic memory in episodic encoding (Tse et al., 2007; Gliga, Volein, & Csibra, 2010), and by building and testing detailed computational models that triangulate between the multiple factors that influence memory performance (Shiffrin & Steyvers, 1997).

Acknowledgements 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, 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, UK: Cambridge University Press.

Craik, F., & Lockhart, R. (1972). Levels of processing: A framework for memory research. levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671-684.

Gliga, T., Volein, A., & Csibra, G. (2010). Verbal labels modulate perceptual object processing in 1-year-old children. *Journal of Cognitive Neuroscience*, *22*(12), 2781-2789.

Goldstone, R., & Hendrickson, A. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(1), 69-78.

Harnad, S. (Ed.). (1987). *Categorical perception: The groundwork of cognition.* New York: Cambridge University Press.

Lupyan, G. (2008). From chair to "chair:" a representational shift account of object labeling effects on memory. *Journal of Experimental Psychology: General*, *137*(2), 348-369.

Sakamoto, Y., & Love, B. (2006). Vancouver, toronto, montreal, austin: Enhanced oddball memory through differentiation, not isolation. *Psychonomic Bulletin and Review*, *13*, 474-479.

Schooler, J., & Engstler-Schooler, T. (1990). Verbal overshadowing of visual memories: some things are better left unsaid. *Cognitive Psychology*, *22*, 36-71.

Shiffrin, R., & Steyvers, M. (1997). A model of recognition memory: Rem - retrieving effectively from memory. *Psychonomic Bulletin and Review*, *4*(2), 145-166.

Sloutsky, V., & Fisher, A. (2004). When development and learning decrease memory. *Psychological Science*, *15*(8), 553-558.

Symons, C., & Johnson, B. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin*, *121*, 371-394.

Tse, D., Langston, R., Kakeyama, M., Bethus, I., Spooner, P., Wood, E., Witter, M., & Morris, R. (2007). Schemas and memory consolidation. *Science*, *316*, 76-82.

Tulving, E., & Thomson, D. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *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. i. on the effect of group formations on the memory trace]. *Psychologische Forschung*, *18*, 299-342.

Winawer, J., Witthoft, N., Frank, M., Wu, L., & Wade, A. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, *104*, 7780-7785.