Journal of Experimental Psychology: General
2008, Vol. 137, No. 2, 348–369

0096-3445/08/$12.00 DOI: 10.1037/0096-3445.137.2.348

From Chair to “Chair”: A Representational Shift Account of Object Labeling Effects on Memory

Gary Lupyan
Carnegie Mellon University and Center for the Neural Basis of Cognition

What are the consequences of calling things by their names? Six experiments investigated how classifying familiar objects with basic-level names (chairs, tables, and lamps) affected recognition memory. Memory was found to be worse for items that were overtly classified with the category name—as reflected by lower hit rates—compared with items that were not overtly classified. This effect of labeling on subsequent recall is explained in terms of a representational shift account, with labeling causing a distortion in dimensions most reliably associated with the category label. Consistent with this account, effects of labeling were strongly mediated by typicality and ambiguity of the labeled items, with typical and unambiguous items most affected by labeling. Follow-up experiments showed that this effect cannot be explained solely by differences in initial encoding, further suggesting that labeling a familiar image distorts its encoded representation. This account suggests a possible mechanism for the verbal overshadowing effect (J. W. Schooler & T. Y. Engstler-Schooler, 1990).

Keywords: categorization, labels, recognition memory, representational change, overshadowing

Supplemental materials: http://dx.doi.org/10.1037/0096-3445.137.2.348.supp

> When we speak of “table” we do not mean a special given table with all the accidental properties, but we mean “table” in general. [We] employ the word “table” in this categorical sense when naming a particular table. (Goldstein, 1936/1971, p. 349)

Using words to communicate is a human universal. It is also unique to our species. As far as we know, no other species in its natural environment habitually learns associations between things out in the world and arbitrary signs. It is thus important to try to understand how this universal human practice affects cognitive processing. One notable property of words is that they denote entire categories (e.g., Goldstein, 1936/1971; Harnad, 2005, for discussion). The words *dog*, *bark*, and *jump* do not refer to a particular dog or a given act of barking or jumping but denote categories of objects, sounds, and types of action, respectively. The aim of the present work was to investigate the consequences of labeling familiar items with their category (i.e., basic level) names. This aim is separate from the intuitive idea that the *meanings* communicated by words affect how we represent and remember objects. For instance, in the classic study of Carmichael, Hogan, and Walters (1932) participants viewed ambiguous figures and were subsequently asked to redraw them. After viewing a figure resembling an *X*, participants redrew the figure differently depending on whether it was called a table or an hourglass. Clearly, object names are a source of information and thereby contribute to one’s interpretation of otherwise ambiguous objects (though this particular effect appears to be produced during recollection rather than during encoding; Hanawalt & Demarest, 1939). The current work takes as its aim understanding whether names generated by the participants themselves augment representations of familiar and well-determined objects. On viewing an object known to be a chair, what happens when one classifies it as a “chair” by labeling it? What is the consequence of using a name to classify familiar objects?

One plausible answer is that there is no consequence. If natural language is just a medium for the expression of knowledge and does not alter the form or content of representations, naming things ought not to affect their representations (Li & Gleitman, 2002; for a discussion, see Carruthers, 2002; Fodor, 1975). Even if one accepts that words play a role in shaping human categories by selecting the few relevant ones out of all those that are possible (Gleitman & Papafragou, 2005), it does not necessarily follow that once words are learned they have any effect on representations (Bloom, 2001; Bloom & Keil, 2001). An alternative position is that labels have an effect on representations of category members by highlighting the relationship between items and their categories. Because labels denote entire categories, naming a particular chair with the category label “chair” might alter the competition between bottom-up and top-down sources of activation, resulting in a representation of a particular chair that is more influenced by previously encountered category members.

Knowing what an object is and calling it by its name are in principle separate processes. Indeed, there is evidence that an object’s semantics and its name constitute dissociable knowledge.

Gary Lupyan, Department of Psychology, Carnegie Mellon University and Center for the Neural Basis of Cognition, Pittsburgh, PA.

I thank Jay McClelland, David Rakison, Brian MacWhinney, Anna Fisher, Dwight Kravitz, and Larry Barsalou for useful commentary and insightful suggestions, and Brian Mathias, Ashleigh Molz, Judith Vogel, Rebecca Snyder, and Karen Doersch for help with data collection. This work was in part supported by a National Science Foundation Graduate Student Fellowship.

Correspondence concerning this article should be addressed to Gary Lupyan, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213-3890. E-mail: glupyan@cnbc.cmu.edu

For instance, Druks and Shallice (2000) described a patient who, when presented with a picture of a kangaroo, was able to provide a detailed encyclopedic description for it but was unable to label the image. It is certainly possible to categorize without naming—this happens every time we recognize something but do not know its name (Harnad, 2005)—and even pigeons have been argued to form categories (e.g., see Astley & Wasserman, 1992, 1998, for pigeons responding categorically to cars, flowers, etc.). But it is only humans that have names for their categories. Whereas categorization of familiar items is rapid and arguably automatic (Grill-Spector & Kanwisher, 2005), an overt classification response using the object's category name may further augment the representation of the item with top-down category information associated with the label. In this view, an object's name is more than just the output of the conceptual system (see Gleitman & Papafragou, 2005, for discussion) but rather can feed back to alter the representation of the labeled item (Clark & Karmiloff-Smith, 1993; Dennett, 1994; Pederson et al., 1998).

Although few studies have investigated the effects of labels on representations of familiar items and categories, the degree to which verbal labels shape the learning of categories, particularly of infants and children, has been extensively studied. One of the most basic findings is that from a very young age, words draw our attention to object categories. Nine-month-old infants, for example, pay more attention to labeled than to unlabeled objects (Balaban & Waxman, 1997), and contrasting words (e.g., *duck* and *ball*) can facilitate object individuation in infants (Xu, 2002). Later in development, calling things by the same name leads children to look for similarities among objects (Loewenstein & Gentner, 1998; Smith, Jones, & Landau, 1996; Waxman & Markow, 1995), whereas calling things by different names leads children to treat the objects as more distinct (Katz, 1963; Landau & Shipley, 2001). In addition, labels, acting as cues to categories, facilitate inductive inferences in children (Gelman & Markman, 1986), possibly by competing with perceptual similarity (Sloutsky & Fisher, 2004a), and promote taxonomic over thematic groupings (Markman & Hutchinson, 1984; Waxman & Hall, 1993).

There are also effects of labels in the learning of novel categories. Lupyan, Rakison, and McClelland (2007) showed that labeled categories were learned faster than unlabeled categories even though the labels did not bring any additional information to the task and participants could not rely on the labels to perform the categorization task. It is also known that associating unfamiliar objects with additional semantic information—"this one is sleepy and angry; this one is fast and lazy"—can actually alter visual processing in adults, facilitating perceptual identity judgments for items with overlapping conceptual associates (Gauthier, James, Curby, & Tarr, 2003). The mechanisms by which labels exert these effects remain elusive. The present work is an effort to formulate a theoretical basis for understanding how names affect cognitive processing within a single domain: recognition memory.

Within what is arguably the most influential class of single-item recognition models, performance is based on the outcome of a global-matching retrieval process (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). Correctly identifying previously seen items depends on a match between the retrieval cue and the contents of memory. The closer the match between the two, the more familiar the item seems and the more likely it is to be correctly recognized. Manipulations such as repeated presentations of a study item and longer presentation times are predicted to improve recognition accuracy through an increase in the number or fidelity of stored features—the more stored features, the greater the similarity between the memory contents and the retrieval cue (Gronlund & Ratcliff, 1989). Conversely, manipulations that decrease the match between memory contents and the retrieval cue are predicted to decrease recognition accuracy.

In the present work, it is hypothesized that when category labels are activated, they produce top-down feedback that activates visual features stored with the category on previous occasions. The features activated by top-down processing become coactive with features activated through bottom-up processing. As activation patterns continue to cycle, the active visual features settle on those that are consistent with both bottom-up input from the exemplar and top-down input from the category (McClelland & Rumelhart, 1981). This results in a mismatch between the stored representation of the studied item and the retrieval cue (the studied item presented during the testing phase), which in turn should produce more "new" responses for old items (i.e., a lower hit rate). When category labels are less active, or when the top-down activity is interfered with, the representation of the visual input is closer to the bottom-up information presented than when labels produce top-down inferences about the category's features. This results in a closer match between the studied item and the item presented at test (the retrieval cue), thus resulting in a higher proportion of hits. In summary, hit rates should be lower for overtly classified (labeled) familiar items than for ones that are not overtly classified. This proposed mechanism is referred to as the *representational shift account*.

The idea of higher level information augmenting lower level representations through top-down feedback is familiar in vision research. Gauthier et al. (2003), for instance, have found that learned conceptual associations affect the speed and accuracy of same–different judgments. Using stereoscopic depth cues to manipulate perception of three-dimensional surfaces, He and Nakayama (1992) have shown visual search to be strongly affected by perceptual completion, and Suzuki and Cavanagh (1995) have demonstrated that the engagement of high-level features (facial expressions) produces perceptual grouping that makes it more difficult to isolate single features from schematic faces. All of these results depend on a top-down modulation of lower level representations by higher level ones.

Although no previous work has directly tested the prediction that producing a labeling response should result in poorer recognition memory, there is evidence that study contexts that promote a kind of categorical coding of items result in reduced memory compared with study contexts that encourage more item-specific encoding. For instance, Marks (1991) showed that memory for pictorial details is enhanced by judging the physical features of pictures compared with judging whether pictures fit into scenes, a response requiring semantic processing of the items. Koutstaal and Schacter (1997) found that adults, and in particular older adults, showed substantial false recognition of novel detailed color pictures from studied categories. For instance, after studying several pictures of cats, participants failed to reject as novel a new picture of a cat. The authors hypothesized that perhaps after participants had seen several cats, the features of a new cat were increasingly likely to resemble features that had been encountered in items presented earlier. A similar finding was obtained by Sloutsky and

Fisher (2004b), who compared the performance of 5-year-old children with that of young adults in their ability to recognize previously seen items and to reject new items from the studied categories. They found that whereas adults had higher accuracy under normal study conditions, children outperformed adults in their ability to correctly reject novel items when the study phase involved category induction—a context promoting category-level encoding in adults but apparently not in 5-year-old children. Thus, in contexts that promote encoding at a coarser level, whether gist, category, or schema, memory for item details appears to be compromised. The hallmark of such "coarse" encoding is a higher false alarm rate for items from the studied category, or items closely associated with the studied ones (e.g., Roediger & McDermott, 1995)—effects that have been shown to arise in connectionist systems using distributed representations (McClelland, 1995).

The false-recognition studies cited above and the representational shift account both predict poorer recognition memory for labeled items but for different reasons, with different patterns of results. False recognition refers to falsely recognizing novel items as old, as reflected in higher false alarm rates. So, one may predict that insofar as labeling a familiar picture would lead to a category-level encoding of the items, it might also result in high false alarms owing, perhaps, to encoding only the features relevant for the category. False recognition ensues because these features are shared by many category members (Koutstaal & Schacter, 1997). The representational shift account, in contrast, predicts lower *hit rates* for the labeled items owing to the stored representations failing to match test-item representations. This is a post-encoding effect. That is, all items are predicted to be initially encoded in a similar way, but representations of the labeled items are distorted by the top-down modulation from the category label.

The representational shift account also predicts that the effect of classification on memory should be mediated by the typicality of the item being overtly classified (i.e., labeled). Not all items are equally good members of a given category. Although the label "chair" is discrete, it is more strongly associated with some chairs than with others, as evidenced by the relative ease of labeling typical compared with atypical items (Rosch, 1978). Overall, typical objects are more easily accessed than atypical objects (Kail & Nippold, 1984; Rosch, 1973), a finding that is easily extended to labeling because most measures of access are measures of naming or labeling latency (Rosch, 1978). One prediction may be that typical items, already being tightly linked to the category, will be less affected by labeling—a typical chair is already a good example of the category, and so further highlighting the category by using a label may not have much effect on its representation. Alternatively, and perhaps less intuitively, typical and unambiguous items may be more affected by labeling. As category labels become activated more strongly, they contribute greater top-down activation to the visual representation of the object than when they are less strongly activated. Because typical items activate category labels more strongly than atypical items, they would distort visual representations to a greater degree, resulting in a greater study-to-test mismatch and thus poorer recognition accuracy.

Unlike a manipulation that involves assigning meaningful labels to ambiguous or abstract stimuli (Carmichael, Hogan, & Walters, 1932; Koutstaal et al., 2003; Musen, 1991), the present experiments made use of highly familiar stimuli that had clear preexisting semantics. The critical manipulation was of the desired response during the study phase. Participants were asked to classify some items (i.e., respond with the items' basic-level names), and for other items, to make a decision independent of the objects' category (a preference judgment). Although there is little doubt that participants are likely to implicitly categorize all objects (e.g., Grill-Spector & Kanwisher, 2005), asking participants to respond with a category-irrelevant response is hypothesized to attenuate the effects of the label as compared with making an overt labeling response. Whatever the effect of labeling, it should be stronger when participants are asked to label a stimulus compared with when they are asked to make a response independent of the label.

Experiment 1 was designed to test whether naming familiar objects (chairs and lamps) produces a decrease in recognition accuracy of the labeled objects. A second goal was to examine whether the effect of naming on memory depended on the typicality of the classified items. If naming reduces recognition accuracy, does it reduce it more for typical or less typical items? Experiment 2 used categories that were more similar to each other—chairs and tables—to examine whether effects of naming on memory are mediated by how ambiguous a labeled item is with respect to the category. Experiments 3–5 were designed to rule out alternative explanations based on levels of processing and differences in encoding strategies between the study conditions. Finally, Experiment 6 tested a prediction arising from the representational shift account: that items are perceived as more typical in the presence of category labels.

Experiment 1

Method

Participants

Eighteen participants (mean age = 21.7 years, *SD* = 4.8 years) gave informed consent and received course credit or $7, in compliance with the Institutional Review Board of Carnegie Mellon University.

Materials

Forty pictures of chairs and 40 pictures of lamps were selected from the IKEA online catalog (www.ikea.com). Each picture showed a single chair or lamp on a white background (see Figure 1 for examples) surrounded by a uniform black background. The stimuli were 250 × 250–pixel color images presented on a 17-in. CRT monitor. The participants viewed the images at a distance of approximately 72 cm, with images centered on the screen and subtending about 6° of visual angle. Participants made responses using a gamepad controller. During the study session, participants used a total of four separate buttons. Two buttons were used for the category responses (chair vs. lamp), and two different buttons for the preference judgments (like vs. don't like). During the test session, participants used two buttons: one to respond *old* and another to respond *new*. Stimulus presentation was controlled by Presentation software (Version 9.20; www.neuro-bs.com).

The full experimental stimulus set comprised 20 chairs and 20 lamps used during study and 20 new chairs and 20 new lamps used as lures at test. The stimuli were selected at random from a larger set with the stipulation that for each picture, a matched *critical lure* was also selected. Lures differed from the studied items in small

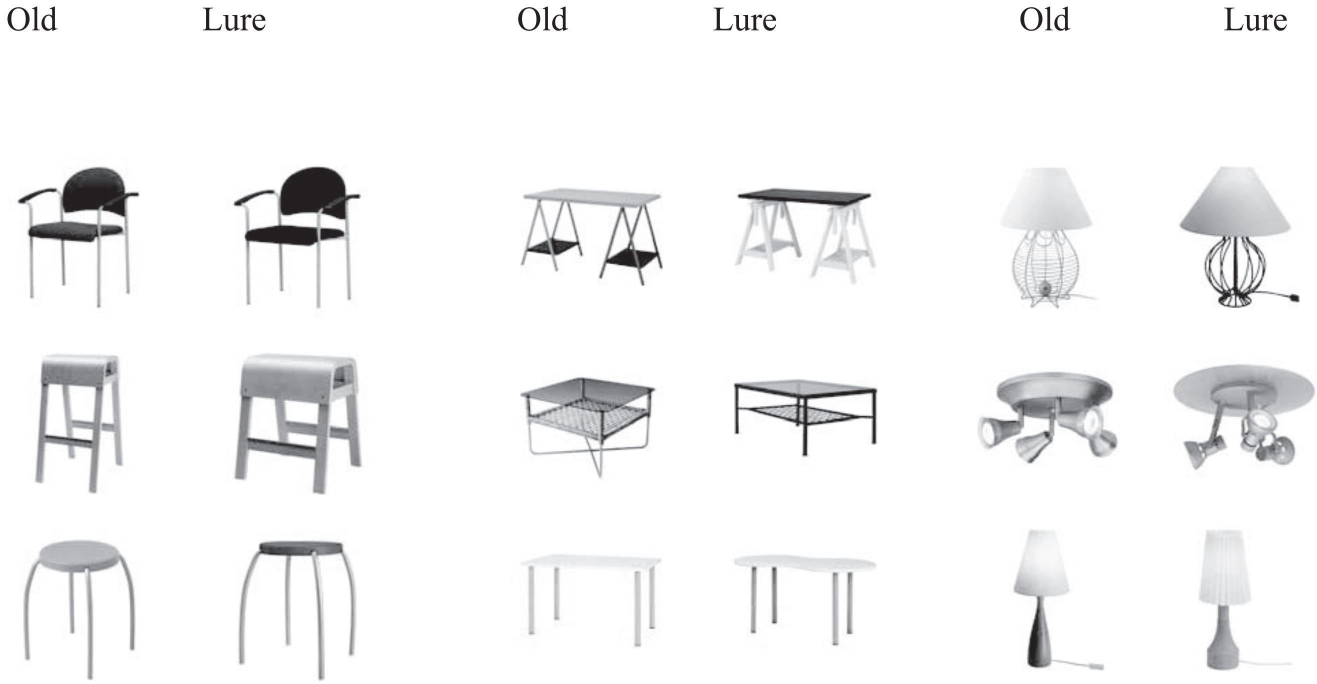

Figure 1. Sample study items and lures for Experiments 1, 3, 4, 5, and 6 (chairs and lamps) and Experiment 2 (chairs and tables). The first pair of chairs differ in color only—see a color version of this figure on the Web at http://dx.doi.org/10.1037/0096-3445.137.2.348.supp

but noticeable ways: The lure might be a different color from the studied item, differ on a feature such as armrests, or be a slightly different shape—for example, a narrower lamp versus wider lamp (see Figure 1). Matching lures to specific items made it possible to compute separate false alarm rates for the two conditions despite using a within-subject design.

Procedure for Collecting Typicality Ratings

To investigate the effects of typicality on naming and memory, typicality ratings for chairs, lamps, and tables (for Experiment 2) were collected from 10 participants (mean age = 19.1 years; *SD* = 1.1 years) who did not take part in any of the other experiments. The items were presented one at a time, with all the items of one category presented before proceeding on to the next. Each item was presented twice. Category order was counterbalanced. For each item, participants responded to the question “How typical is this [chair/table/lamp]?,” rating each item on a 5-point scale ranging from 1 (*very typical*) to 5 (*very atypical*). Each picture remained on the screen until a responses was made.

Experimental Procedure

Study phase. The experiment consisted of two phases: study and test. Participants were instructed that they would see a number of pictures of chairs and lamps, with half of the pictures presented in *classification* blocks and half in *preference* blocks. Participants were told that for the classification blocks, they should classify each picture as a chair or a lamp by responding with one button for chairs and another for lamps. For the preference blocks, participants were told to indicate their preference for the objects displayed in the pictures: one button for *like*, another for *don’t like*. Participants were also told that the pictures would be presented quickly and that they must pay careful attention to each one and try to remember as much as possible about each picture.

Of the 80 total pictures, 40 were used in the study phase. The remaining 40 pictures were reserved for the test phase, as matched lures (see Figure 1). The study phase consisted of eight blocks of 10 trials each. Although participants were asked to remember as much as possible about each item, they were not explicitly told there would be a memory test. Before each block, participants saw instructions indicating whether they should perform classification or preference judgments after each trial. The conditions alternated—classification, preference, classification, and so on. To increase overall recognition memory, each stimulus was presented twice. For instance, a given table might be seen in Blocks 1 and 5 or Blocks 2 and 6. The condition of the starting block was counterbalanced between participants. A particular participant saw a given item—say, Chair A—in either a classification or a preference context. However, each participant saw 20 chairs and lamps in a preference context and the remaining 20 in a classification context, allowing for within-subject comparisons of recognition memory. Across the group of participants, every item was seen in both contexts, allowing for between-subjects item analysis.

Each study trial began with a presentation of a fixation cross for 1,500 ms, followed by a randomly selected picture for 300 ms. After the picture disappeared, a response cue consisting of a question mark appeared for 700 ms, during which time participants had to make a response (*chair*/*table* or *like*/*don’t like*, depending on the current block). The labels "chair" and "table" were included in the initial instructions, but the actual response cue consisted of the question mark only. Participants were instructed to respond only while the response cue was on the screen. The relatively short stimulus presentation times and response windows served to limit the degree to which participants "thought" of the objects in terms of the verbal labels when making the preference judgments. Though in all likelihood participants knew the category of the objects for which they indicated preference, limiting the time to consider the category label was done to enhance the difference between the classification and preference conditions.

To familiarize them with the pace of the experiment, prior to starting the study trials, participants viewed five pictures of natural scenes shown at the same rate as the actual study images.

Test phase. After completing the study phase, participants were told that they would now see more pictures; some would be exactly the same as those seen before, whereas others would be new but similar—differing subtly in details like shape or color. For each picture, the participants' task was to indicate whether they had seen the exact picture before by responding *old* or *new*. This design produced four kinds of stimuli: old pictures seen in the classification context, old pictures seen in the preference context, lures most similar to pictures seen in the classification context, and lures most similar to pictures seen in the preference context. Unlike the study phase, in which classification and preference contexts were blocked, in the test phase all pictures were intermixed. Each picture remained on the screen until a response was made. The test phase ended when participants responded to all 80 pictures.

Results

Typicality Ratings

There were no differences in the rated typicality between the item categories or between old and new items, all $F(1, 38) < 1$. Mean typicality values for the chairs and lamps were, respectively, 2.76 ($SD = 0.94$) and 2.58 ($SD = 0.97$).

Study Phase

Participants classified the objects with an overall accuracy of .90 ($SD = .30$). Classification accuracy is defined here as placing the object into the category intended by the manufacturer of the object, and so it is admittedly subjective. Classification errors were likely inflated by both the short stimulus presentation time and the short response window, as well as by the categorical vagueness of two chairs that some participants classified as lamps. Invalid responses—classifying in a preference block or indicating preference in a classification block—accounted for about 1% of the trials and were marked as incorrect responses. Responses made outside the response window were omitted from the study phase analyses.

More typical items were classified more quickly and more accurately: There were significant correlations between rated typicality and labeling accuracy, Pearson $r(38) = -.36$, $p = .02$, and between typicality and classification response time (RT), $r(38) = .55$, $p < .0005$. The correlation between typicality and preference judgment times was not significant, $r(38) = -.14$, $p > .3$. For preference judgments, 50.3% of the valid responses were *like* and 49.7% were *don't like*. More typical items were better liked, $r(38) = .46$, $p < .01$. Likability of an item did not correlate with any RTs or recognition memory in any of the experiments and is not discussed further (the one exception was Experiment 5, where greater preference in the study phase predicted marginally higher hit rates in the subsequent test phase, $p = .09$).

No difference in accuracy or in hit rates or false alarms was found between participants performing the classification block first and those performing the preference block first: one-way analysis of variance, $F(1, 16) < 1$; the reported results therefore collapsed across all participants. There were no differences between hit rates or false alarms for chairs and lamps, $F(1, 16) < 1$.

Test Phase

Despite seeing the study items for only 300 ms, participants' recognition memory was considerably above chance (see Table 1). Both the hits, $t(17) = 14.33$, $p < .0005$, and the false alarms, $t(17) = 6.62$, $p < .0005$, were significantly different from the chance value of .5. Participants had lower recognition memory for items they had classified compared with those for which they had indicated preference, as measured by accuracy (hits minus false alarms; pairwise t test), $t(17) = 4.53$, $p < .0005$, as well as d', $t(17) = 4.62$, $p < .0005$. The difference in recognition memory arose from a difference in hits. Participants had a significantly lower hit rate for items they had classified with the category name compared with items for which they had indicated preference, $t(17) = 6.13$, $p < .0005$. To compute effect size, Cohen's d was computed using Hedges's adjustment for sample size: $d = 2t / \sqrt{(df)}$, resulting in an effect size of 2.04, which constitutes a very large effect. Because each item was matched to a critical lure, it was possible to calculate separate false alarm rates for the two conditions. Although participants saw the lures for the first time during test, each lure was most similar to an object previously seen in either a classification or a preference block. There was no significant difference between the lures most similar to the classified items and the lures most similar to the items for which preference had been indicated, $t(17) < 1$. A summary of the hits

Table 1
Mean Proportion (and Standard Deviation) of Hits and False Alarms, Mean Accuracy (Hits Minus False Alarms), and d' for Experiments 1–2

Experiment and condition	Hits	False alarms	Accuracy	d'	Effect size of d' difference between conditions (Cohen's d)
Experiment 1	*		*	*	
Preference	.83 (.09)	.34 (.14)	.49	1.49	1.17
Classification	.62 (.10)	.30 (.15)	.32	0.90	
Experiment 2	*		*	*	
Preference	.82 (.10)	.36 (.18)	.46	1.41	0.69
Classification	.63 (.15)	.30 (.12)	.33	0.94	
Experiment 3	*	*	*	*	
Preference	.75 (.10)	.33 (.14)	.42	1.20	1.04
Classification	.65 (.17)	.42 (.13)	.23	0.63	

* $p = < .05$.

and false alarm rates, accuracy, and d' for Experiments 1–3 is presented in Table 1.

Typicality Effects at Test

To test whether the effect of labels on memory was predicted by typicality of the study items, previously collected typicality ratings were correlated with recognition memory for the study items when they were studied in the classification versus preference blocks. Overall, participants had a higher hit rate for atypical items than for typical items, $r(38) = .64$, $p < .0005$. When typicality was entered as a covariate in a general linear model to predict hit rates, the analysis revealed a significant typicality by study condition interaction, $F(1, 76) = 7.82$, $p < .01$. This shows that the relationship between hit rates and typicality was significantly stronger for the classified items compared with the items for which preference was indicated.

False alarm rates correlated marginally with typicality, $r(38) = –.30$, $p = .06$—more typical items had slightly higher false alarms. Using typicality as a covariate in a general linear model to predict false alarms failed to reveal a significant typicality by study condition interaction, $F(1, 76) < 1$.

To summarize, the difference in hits between the classification and preference conditions was largest for the typical items. The difference in false alarms between the two conditions did not vary as a function of typicality.

Discussion

Experiment 1 confirmed the prediction that labeling familiar items with their basic-level names results in poorer subsequent recall. The difference in memory between the two contexts (classification and preference) was reflected as a difference in hits rather than false alarms. Finally, the effect of labels was strongly mediated by typicality. Atypical items were remembered well regardless of what context they were in. Memory for the more typical items, however, depended strongly on study context, with labeling producing the largest decrement in hits for the most typical items. This may seem strange when one considers that typical items are already similar to the category prototype and so have less potential to be affected by top-down feedback from the category representation than atypical items. Although it is true that a typical item, by definition, is more similar to the category prototype, it may be argued that a true prototype of a complex object (as compared with, say, colors or simple shapes) is never encountered. Thus any real object, however typical it may seem, can always "drift" closer to the theoretical prototype (e.g., see Experiment 6). Thus, although typical objects are in fact more similar to the category prototype than atypical objects, the degree of representational shift may depend more on the strength of the attractive force of the category on the exemplar, which is stronger for typical items and weaker for atypical ones. This issue is addressed in more detail below (see especially *Shift-to-Prototype and Perceptual Magnet Effects*).

It may also appear anomalous that labeling resulted in lower hits rather than higher false alarms. If naming resulted in coarser, more category-based encoding, one would expect participants to err by having higher false alarms (e.g., Koutstaal & Schacter, 1997; Sloutsky & Fisher, 2004b). A difference in hits, however, is exactly what would be expected if recognition accuracy depends on a close match between a memory representation and the retrieval cue (the original study item re-presented at test) (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). The study-to-test mismatch due to naming occurs if naming shifts the representation of the labeled item away from the retrieval cue. If this is the case, finding that memory is affected most for the typical items indicates that naming these items produces the largest representational shift.

Experiment 2

Experiment 1 demonstrated that hit rates are markedly lower for the classified items and that the impairment is conditioned by typicality. Experiment 2 was designed to replicate this effect with categories that had a greater perceptual overlap: chairs and tables. Using these categories made it possible to investigate the effect of category ambiguity on memory. Recall that Carmichael et al. (1932) found distortions in reproductions of ambiguous items named by the experimenter. Labeling ambiguous items may have induced the participants to augment the visual representation with conceptual (i.e., categorical) representations, which acted to distort the memory. In this view, ambiguous items should be misremembered more after labeling than unambiguous items. The representational shift account predicts an opposite pattern of results: Because ambiguous items have a weaker association with the category labels compared with unambiguous items, labeling them should produce less distortion than labeling unambiguous items. Thus, recognition of ambiguous items should not be impaired by classification, in part because these items have weaker links to the category label, thereby resulting in less effective top-down modulation of their visual representations. Experiment 2 tested this prediction by correlating ambiguity and typicality ratings with recognition memory.

Method

Participants

Eighteen participants (mean age = 18.9 years, $SD = 1.3$ years) gave informed consent and received course credit.

Materials

The materials were identical to those used in Experiment 1 except pictures of lamps were replaced with pictures of tables, also obtained from the IKEA online catalog.

Procedure for Collecting Ambiguity Ratings

A separate group of 18 participants (mean age = 20.4 years, $SD = 2.1$ years) contributed category ambiguity (i.e., category vagueness) ratings for the chairs and tables. Images of chairs and tables were presented in random order together with the question "Is this a chair?" or "Is this a table?" Each picture was presented twice, once in each context. The choices were *yes*, *no*, and *uncertain* (Hampton, 2006). Ambiguity was computed as the difference between proportions of *yes* and *no* responses to each item, averaged across context. For instance, an item that was given 90% *yes* and 5% *no* responses to the question "Is this a chair" (with the remaining 5% comprising "uncertain" responses) and 88% *no* and 5% *yes* responses to the query "Is this a table" would have an ambiguity rating of $[(90 – 5) + (88 – 5)] / 2 = 84$. The most unambiguous chair would have a value of 100 (100% of responses are *yes*). The most ambiguous chair would have a value of 0 (e.g., 50% answering *yes* and 50% answering *no*).

Results

Ambiguity Ratings

Figure 2 shows the relationship between rated typicality and ambiguity, $r(38) = -.42$, $p < .01$ (top) and the distributions of the ratings (bottom). Mean ambiguity values for chairs and tables, respectively, were 75.28 (SD = 22.69) and 72.36 (SD = 17.67).

Study Phase

The first aim was to establish that the chair–table categories were indeed more confusable than the chair–lamp categories. Overall classification accuracy was 83.1% (SD = .16), which was significantly different from classification accuracy in Experiment

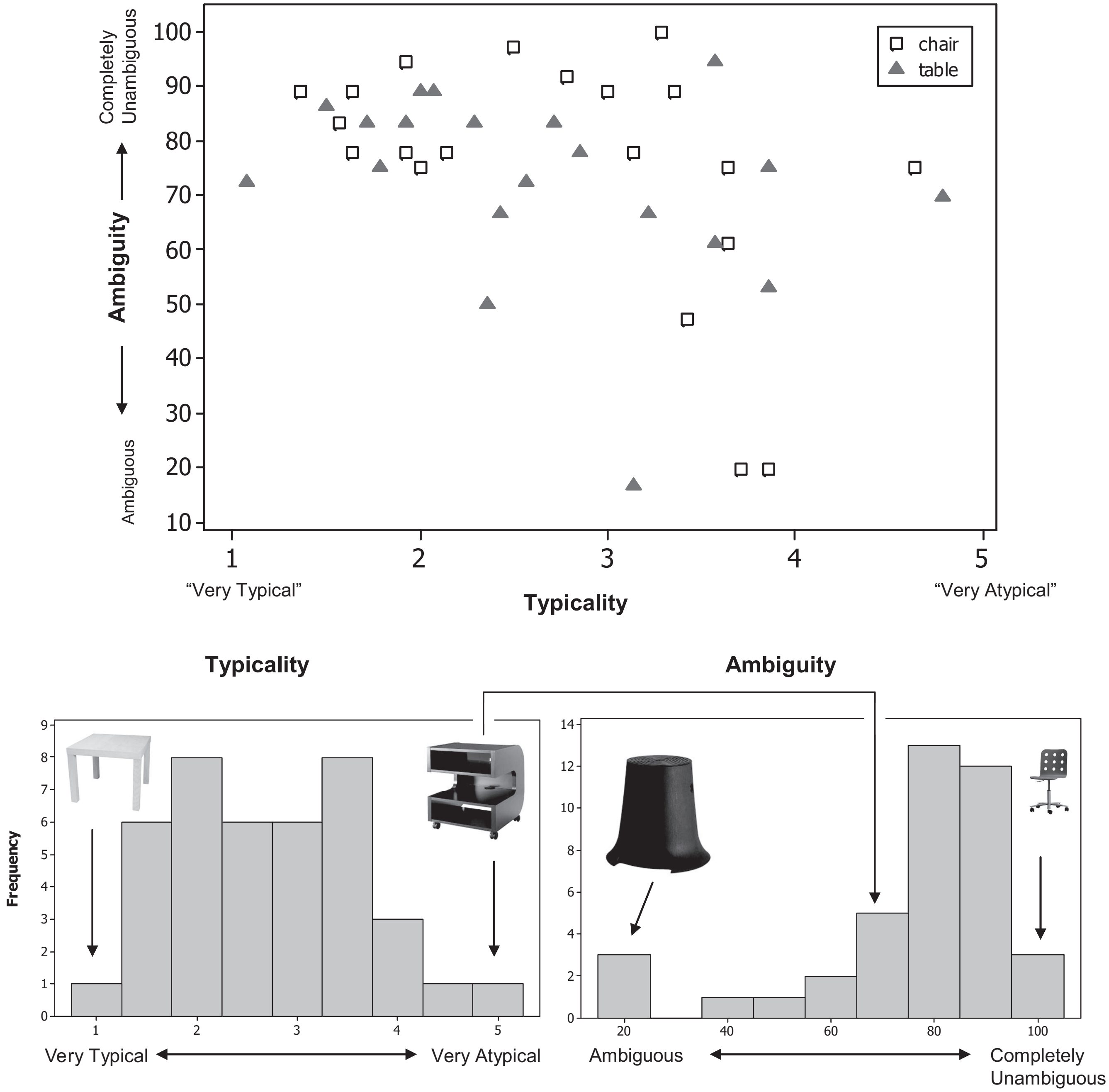

Figure 2. The relationship between typicality and ambiguity ratings for the chair and table items used in Experiment 2 (top) and distributions of the typicality and ambiguity ratings (bottom), along with representative examples. Note that items rated as being most atypical are not the most ambiguous.

1 ($M = 90.0\%$), t test of all items, $t(78) = 2.41$, $p < .025$. Further evidence of greater confusability between chairs and tables compared to chairs and lamps comes from a difference in classification RTs, with longer RTs in the present experiment ($M = 357$ ms, $SD = 124$ ms) compared with Experiment 1 ($M = 301$ ms, $SD = 137$ ms), $t(78) = 3.86$, $p < .0005$. Restricting the analysis to the chairs only, classification accuracy in Experiments 1 and 2 was not reliably different—87.8% versus 83.5% ($t < 1$). However, RTs for classifying chairs among tables were marginally greater, as indicated by both a pairwise item analysis, $t(19) = 2.06$, $p = .06$, and an analysis of response RTs by subjects (for chairs among tables, $M = 341$ ms, $SD = 44$ ms; for chairs among lamps, $M = 304$, $SD = 56$ ms), $t(17) = 2.21$, $p < .05$. Together these analyses suggest that the chair–table category distinction was more difficult than the chair–lamp distinction, most likely owing to greater perceptual similarity between the chair–table categories.

Participants classified more typical items both faster, $r(38) = .55$, $p < .0005$, and more accurately, $r(38) = -.44$, $p < .005$, than less typical items. The very same relationship held for ambiguity, with unambiguous items classified more quickly and accurately. When typicality and ambiguity were both entered into a multiple regression, only ambiguity predicted classification accuracy: $\beta_{\text{ambiguity}} = .005$, $t(39) = 4.62$, $p < .0005$, $R^2 = .46$. Both typicality and ambiguity were simultaneously significant predictors of classification RT:$\beta_{\text{typicality}} = .28$, $t(39) = 4.23$, $p < .0005$; $\beta_{\text{ambiguity}} = .80$, $t(39) = -2.68$, $p < .02$; $R^2 = .50$.

Test Phase

Participants had reliably lower hit rates for items they classified, $t(17) = 5.25$, $p < .0005$, $d = 1.72$. There was no significant difference in false alarms; however, a trend was found with participants having marginally lower false alarms for lures most similar to items studied in the classification condition compared with lures most similar to items studied in the preference condition, $t(17) = 1.82$, $p < .09$. As in Experiment 1, there were significant differences in overall recognition performance as measured by accuracy, $t(17) = 3.03$, $p < .01$, and d', $t(17) = 2.54$, $p < .03$. There were no reliable differences in recognition accuracy between classification-first and preference-first conditions, $F(1, 16) < 1$; the reported results collapse across this factor.

Analyses of the Effects of Typicality and Ambiguity on Memory

Experiment 2, like Experiment 1, revealed a relationship between hit rates and typicality. Typicality mediated recognition for the classified items, $r(38) = .54$, $p < .0005$, while having no effect on items for which preference was indicated, $r(38) = .18$, *ns*. This resulted in a significant condition by typicality interaction, $F(1, 76) = 5.22$, $p < .03$ (see Figure 3, left). This interaction is clarified by restricting the analysis to items with typicality ratings in the first and fourth quartiles, that is, the most typical and most atypical items. This condition by typicality interaction was highly significant, $F(1, 38) = 8.60$, $p < .01$ (Figure 3, bottom left).

Ambiguity correlated significantly with $\text{hits}_{\text{classified}}$, $r(38) = -.35$, $p < .025$, but not with $\text{hits}_{\text{preference}}$, $r(38) = -.08$, *ns* (Figure 3, right). The condition by typicality interaction with ambiguity as a covariate did not reach significance, $F(1, 76) = 2.46$, $p = .12$.[1] However, comparing the effect of classification and preference on the most unambiguous and most ambiguous items (those from the first and fourth quartiles of ambiguity ratings) revealed a significant condition by ambiguity interaction, $F(1, 48) = 5.72$, $p < .025$ (Figure 3, bottom right). Classification affected the least ambiguous items more than the most ambiguous items. False alarm rates were not predicted by typicality or ambiguity, with respective correlation values of $r(38) = .12$, *ns*, and $r(38) = .03$, *ns*.

Because overall hit rates were much higher after preference judgments than after classification, it is possible that the stronger effect of typicality/ambiguity on memory after classification was due to a ceiling effect of the high hit rates after preference judgments. If so, the high hit rates in the preference condition would mask a potential effect of typicality and ambiguity on this measure. To address this possibility, items for which the hit rate following preference judgment exceeded the median rate were excluded, and the correlations recomputed. After excluding the best remembered items for which preference was indicated, there was no significant difference in hit rates between the two conditions, $t(19) < 1$. However, hit rates for the classified items were still mediated by typicality, $r(18) = .59$, $p < .01$, and somewhat by ambiguity, $r(18) = -.33$, $p = .15$, whereas hit rates for items in the preference blocks were not predicted by typicality, $r(18) = .02$, *ns*, or ambiguity, $r(18) = .09$, *ns*.

Unsurprisingly, unambiguous items also tend to be more typical than ambiguous items. The two measures were indeed correlated, $r(38) = -.36$, $p < .025$ (Figure 2, top). With ambiguity partialed out, typicality still correlated significantly with $\text{hits}_{\text{classified}}$, $r(38) = .46$, $p < .01$. With typicality partialed out, however, ambiguity no longer predicted $\text{hits}_{\text{classified}}$, $r(38) = -.19$, $p > .20$, suggesting that in part owing to the correlation between typicality and ambiguity, these measures do not make fully independent predictions, with typicality being the more reliable predictor of labeling effects on memory.

Discussion

Participants in Experiment 2 had much poorer memory for the items they overtly classified with the category label compared with those for which they indicated preference. As in Experiment 1, the difference in memory was reflected as a difference in hits rather than false alarms, and the effect was strongly mediated by typicality. The most typical items were more affected by labeling than the less typical items. Other than replicating this main effect, the use of categories with greater perceptual overlap allowed an investigation of how item ambiguity contributed to the effect of labeling found in Experiment 1. It was found that labeling the unambiguous items produced the largest decrement in memory. This result is contrary to the more intuitive prediction that labeling ambiguous items should have the largest effect on memory insofar as it would result in encoding the otherwise indeterminate item in a more categorical fashion.

[1] Strictly speaking, using correlations and F values is inappropriate for the ambiguity ratings because they were not normally distributed (Figure 2). Using nonparametric tests or square-root-transformed ambiguity ratings did not qualitatively change any of the reported results but made it more difficult to compare the relative contributions of typicality and ambiguity.

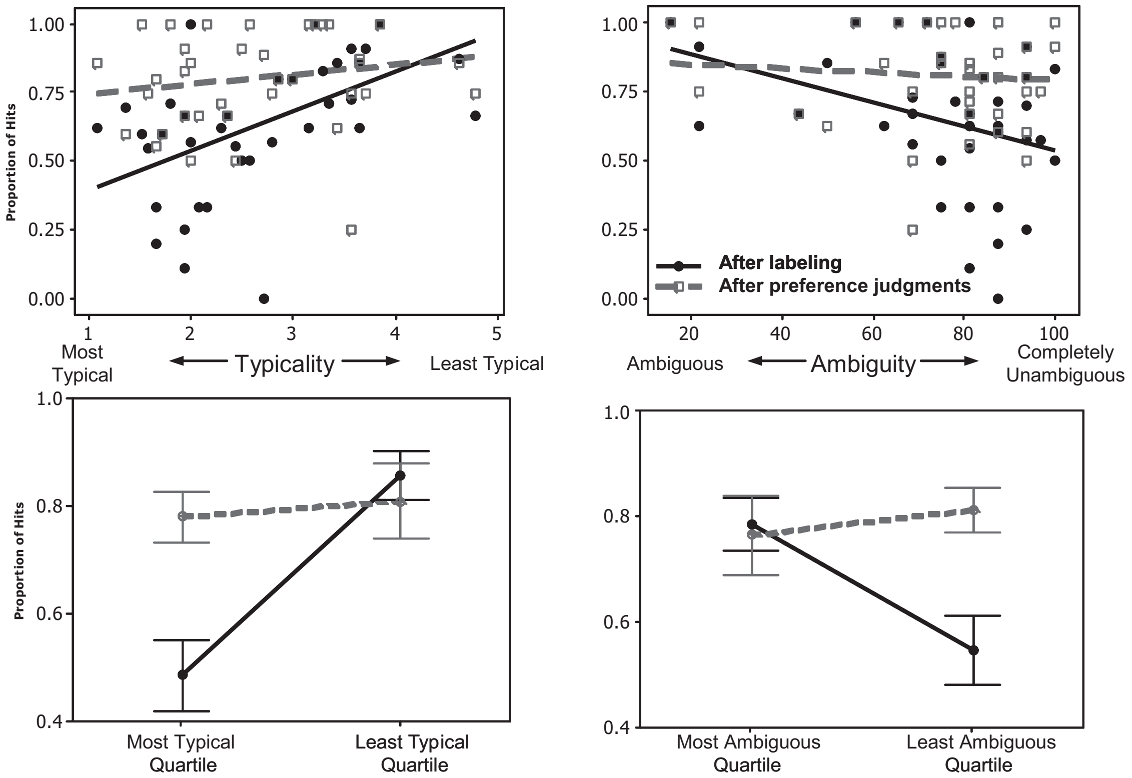

Figure 3. Top: Proportion of hits (±1 *SE*) of chairs and tables as a function of their rated typicality (left) and ambiguity (right) following classification and preference judgments in Experiment 2. Bottom: Average hit rates for the first and fourth quartiles of typicality (left) and ambiguity (right) ratings following classification and preference judgments.

The obtained result is compatible with the representational shift account, according to which it is the unambiguous items that activate the label to the greatest degree and it is these items that tend to have the attributes most strongly associated with category label (i.e., the attributes whose values are most liable to be affected by the top-down feedback from the label).

The failure to find an effect of labels on representations of ambiguous objects may seem to contradict the finding by Carmichael et al. (1932) that labels affect the representations of ambiguous objects: An *X* labeled as a “table” would be redrawn as more tablelike, whereas an *X* labeled as an “hourglass” would be redrawn having more hourglasslike features. In fact, subsequent studies showed that this classic finding was likely due to a reliance on verbal labels in the reproduction process rather than an effect of labels on encoding. Thus, participants were not attempting to simply redraw the original figure on the basis of their memory for what they saw; they were redrawing the table or the hourglass (Hanawalt & Demarest, 1939). Consistent with this explanation, Prentice (1954) found that in a recognition memory task (as opposed to the original drawing task), referring to these ambiguous figures by different labels did not affect performance.

One issue needs further clarification. One may wonder why the item’s ambiguity should be a factor at all given that the task requires overt classification. Once a participant generates and provides a label, shouldn’t there be a top-down effect even if the labeled item is not so clearly in the labeled category? The same question may be asked concerning effects of typicality: Once an item (however atypical) is labeled, shouldn’t the label augment its representation regardless of its typicality? One explanation for the finding that typicality and ambiguity matter is that not only do atypical and ambiguous items activate the category label more slowly and perhaps less strongly (as indicated by longer RTs), but because atypical and ambiguous items tend to possess fewer of the attributes reliably associated with the category, they are less liable to be affected by the category representation. Several examples may help to illustrate this point. Armrests are a feature strongly associated with the category of chairs (although there are many chairs without armrests, having armrests is a good cue that an object is a chair). As the bottom-up information concerning a specific chair becomes augmented by the category information, a typical and unambiguous chair without armrests may be misremembered as having armrests, leading participants to respond *new* when they encounter the original armrest-less chair at test. Labeling is theorized to greatly enhance the degree to which this occurs, leading to the reported memory effects. The features most likely to be affected by the top-down feedback from the category are those that are most predictive of the category. Yet because atypical and/or ambiguous items are likely to lack these features, category labels have little effect on their representations. As a further illustration of the proposed mechanism, consider two relevant attributes of the category “banana”: color and curvature. The present results suggest that labeling a banana may lead one to exaggerate these category-relevant features, remembering a green banana as being more yellow or a mildly curved banana as more curved, as it is integrated into the larger category. It is unlikely, however, that one would remember a (highly atypical) blue banana or a straight banana as being more yellow or more curved, because such an unusual banana would not be integrated into the larger category, whether labeled or not. To use an example from the present stimulus set, one may misremember a dining room chair without armrests as having them, especially when it is labeled with "chair," but one is unlikely to remember a bar stool or beanbag chair as being more chairlike.

Experiment 3

A possible confound in Experiments 1 and 2 was processing time during the study phase. Recall that participants responded to the images only after they disappeared—the exposure time being fixed at 300 ms and the response window fixed at 700 ms after stimulus offset. Nevertheless, it is possible that if classifying was a quicker task than indicating preference, and encoding was terminated by the response, then encoding of the images in the classification blocks may have been more limited than when making a preference judgment. A comparison of RTs for indicating classification and preference in Experiment 1 revealed a significant difference in RTs during the study phase. Classification judgments were performed about 75 ms faster than preference judgments ($M_{\text{classification}} = 300$ ms, $SD = 59$ ms; $M_{\text{preference}} = 378$ ms, $SD = 65$ ms), pairwise t test, $t(17) = 7.28$, $p < .0005$, $d = 1.26$. It is possible that classification resulted in poorer memory because less time was spent encoding the classified images.

Experiment 3 sought to address the possibility that the effect of classification was produced by strategic or encoding differences. For instance, it is possible that participants focused on structural features of the images in the classification blocks and on the more perceptual features, such as color, for the preference trials. Also possible is that encoding a picture in a more conceptual context such as required by a classification response comes at the loss of visual details (Marks, 1991)—though it is not clear why that should be the case. Experiments 1 and 2 also leave open the possibility that worse memory after classification can be explained through a levels-of-processing account (Craik & Lockhart, 1972) if participants processed the items to be classified at a shallower level or encoded them less distinctly. Experiments 3 and 4 address this possibility. In these studies, instead of the study phase being organized into classification and preference blocks, each stimulus was followed randomly by one of two cues—the word *classify* or *preference*—instructing participants which type of response they should make. Participants could not base their perceptual or attentional processing of the images on the desired response, because the response type was unknown until the stimulus had disappeared.

Method

Nineteen participants (mean age = 19.5 years, $SD = 1.3$ years) gave informed consent and received course credit or $7. One participant was eliminated for providing chance-level responses in the study phase. The procedure was identical to that of Experiment 1 with the following exception: Instead of the stimuli being organized into predefined classification and preference blocks, each stimulus was randomly followed by a response cue: the word *classify* or *preference*. Participants were instructed to base their responses on the cue that appeared. On each trial, therefore, participants could make one of four responses—*chair*, *lamp*, *like*, or *don't like*—but depending on the cue, only two of these responses were valid. To accommodate for the extra time necessary to process the cue, the response window was lengthened from 700 ms to 1,200 ms.

Prior to the start of the experiment, participants were given practice in making the responses and getting used to the pace of the experiment by classifying or indicating preference for 10 pictures of jars and plates. In Experiment 3 the experimenter stressed, as before, that participants should pay careful attention to each image and try to remember as much about each one as possible.

Results

Study Phase

Mean study RT was 826 ms ($SD = 75$ ms); this RT was more than twice as long as in Experiment 1, $t(34) = 22.36$, $p < .0005$. Classification responses were still faster than preference judgments, pairwise t test, $t(17) = 8.20$, $p < .0005$. Intermixing the classification and preference conditions seemed to make the study task more difficult. The much longer RTs indicated that participants, not knowing what kind of decision would be asked of them, were less likely to prepare the response beforehand. Participants classified the items correctly 85.0% of the time. This accuracy was somewhat lower than the 90% accuracy observed in Experiment 1, $t(17) = 1.74$, $p = .09$, but arose entirely from the greater number of invalid responses: indicating preference for items cued as classification (*P-to-C* error) and classifying items cued as preference (*C-to-P* error). Whereas such invalid responses were very rare in Experiment 1, accounting for only 1.4% (C-to-P) and 0.4% (P-to-C) for classified and preference items, respectively, in the present experiment invalid responses accounted for 12.0% (C-to-P) and 5.7% (P-to-C) of total responses.

Test Phase

Experiment 3 replicated the pattern of hits observed in Experiments 1 and 2 (see Table 1). Recognition memory (d') was worse for the items followed by the classification cue compared with those followed by a preference cue, $t(17) = 5.92$, $p < .0005$. As before, the classified stimuli had lower hit rates than stimuli for which participants indicated preference, $t(17) = 3.14$, $p < .01$, $d = 0.71$ (see Table 1). In addition to a difference in hit rates, Experiment 3 also revealed a difference in false alarms, with participants having a greater false alarm rate to lures most similar to classified items, $t(17) = 3.05$, $p < .01$, $d = 0.67$. Experiment 3 also revealed a stronger correlation between hit rates and typicality for the classified items, $r(18) = .69$, $p < .001$, than for items for which preference was indicated, $r(18) = .37$, $p = .11$. A comparison between Experiments 1 and 3 of d' averaged across study condition revealed a marginally lower d' in the present experiment, $t(34) = 1.87$, $p = .07$.

Discussion

Experiment 3 replicated the findings of Experiment 1 in an intermixed study context. Memory for the labeled items was poorer even when participants did not know ahead of time which items they should be labeling. Whereas the main effect of labeling on memory in Experiment 1 can be argued to have resulted from differences in encoding, with participants paying attention to and encoding more perceptual features when asked to make preference judgments, this explanation fails to account for the present results. As in Experiments 1 and 2, recognition accuracy was affected most for the items for which the category (chair or lamp) was retrieved.

There are several differences between the results of this experiment and Experiment 1 that are worth noting. First, the study phase RTs were much longer in the present experiment, suggesting that participants were less likely to prepare responses beforehand. Second, the difference between study conditions in the present experiment was numerically smaller than in Experiment 1. Third, unlike Experiments 1 and 2, participants in Experiment 3 showed a significant difference in false alarms, with classified items having lower hits *and* greater false alarms.[2]

The smaller difference in hits between the conditions in this experiment compared with Experiment 1 may have arisen from participants classifying some items cued as preference. If it is the act of classifying that results in poorer memory, then classifying items in the preference condition should produce lower hit rates. Indeed, the hit rates in the preference condition in Experiment 3 were significantly lower than the preference-item hit rates in Experiment 1.

The significant difference between the false alarms in the two study conditions is more difficult to explain, especially considering that the trend in Experiment 2 was in the opposite direction, with classified items having a lower false alarm rate. If classification actually leads to lower false alarms, then the greater number of invalid C-to-P responses compared with P-to-C responses would mean that fewer items followed by the classification cue actually benefited from the lower false alarms brought about by classification.

Although most participants probably based their response on the response cue (*preference* or *classify*) that appeared after stimulus offset, on a poststudy questionnaire, some participants reported thinking of both responses for each item, which is indeed a reasonable strategy. Not knowing what response would be requested, upon seeing a chair, one might label it to oneself as "a chair that I like." The presence of invalid responses further confirms the use of this strategy. It is interesting to note that levels of processing and dual-coding theory accounts (Craik & Lockhart, 1972; Lockhart, Craik, & Jacoby, 1976) predict that encoding an item for multiple types of responses should produce better memory for the item. Not only did the present results not reveal such an improvement, but memory was somewhat poorer than when the study phase was blocked (Experiment 1). Poorer memory, particularly for the preference condition, is expected, however, if participants classified some of the items cued with *preference* and if overt classification impairs subsequent recognition.

Experiment 4

Although participants in Experiment 3 did not know what type of response they would be asked to make until after the stimulus had disappeared, a possibility remains that during the 1,200-ms response window, an iconic memory trace was available to the participants and this memory trace was processed to different degrees in the classification and preference conditions. If this is true, the recognition impairment for the classified stimuli may still be the result of encoding differences rather than a representational shift caused by the category label.

Experiment 4 attempted to limit the perceptual trace through backward masking (Sperling, 1960), thus limiting the degree to which perceptual information was available for encoding after the offset of each study stimulus (e.g., Spencer & Shuntich, 1970). Given the long (by psychophysical standards) 300-ms stimulus duration, there is no reason to suspect that the mask should impair categorization of the stimuli (Enns & Di Lollo, 2000). Rather, if greater recognition memory for the items studied in the preference context arises from greater perceptual encoding from a hypothesized iconic memory store (Coltheart, 1983), and the mask is hypothesized to terminate additional processing of this store (for a review, see Bachmann & Allik, 1976; Enns & Di Lollo, 2000), then masking the study items may remove the advantage for items in the preference condition.

In Experiments 1–3 participants were instructed to try to remember as much as possible about each item in the study phase. Although finding a difference in recognition memory despite such an instruction is an arguably stronger demonstration of the effect of labeling on memory, it is important to demonstrate the generality of the effect in a context without an explicit instruction to memorize the study items. Experiment 4 omitted the instruction to remember the stimuli.

Method

Twenty-two participants (mean age = 19.4 years, SD = 1.5 years) gave informed consent and received course credit for their participation. Two participants were eliminated for having chance-level performance classifying the study items, leaving 20 participants. The procedure was identical to that of Experiment 3 with the following exceptions. First, each picture (appearing for 300 ms, as before) was followed by a mask presented for 300 ms. The mask was a pattern of colorful swirls that occluded the entire picture. Second, the instruction to "try to remember as much as possible about each item" was omitted. Finally, the assignment of items to study and lure groups was now counterbalanced between participants: Half the participants had study and test items identical to those in Experiment 1, while for the other half, the old items and lures were switched.

Results

The inclusion of the mask made responses in the study phase more effortful, as indicated by longer study RTs. Experiment 4 RTs were on average 59 ms longer than Experiment 3 RTs, $t(36) = 2.62$, $p < .01$. The difference in recognition memory between the classification and preference conditions remained in Experiment 4

[2] Experiment 3 was the only experiment to reveal a significant different in false alarms. It was therefore deemed important to try to replicate this effect. A replication (n = 18) failed to find a significant difference in false alarms between the conditions, $t(17) < 1$. In fact, the direction of the effect was reversed ($M_{\text{FA-preference}} = .34$, $SD = .12$; $M_{\text{FA-classification}} = .31$, $SD = .10$). The replicated study again yielded a difference in hits in the predicted direction, $t(17) = 2.74$, $p = .01$

despite masking and the omission of instructions to remember the items. A summary of the hit and false alarm rates, accuracy, and d' for Experiments 4 and 5 is presented in Table 2. Items followed by the preference cue yielded significantly higher d', $t(19) = 2.71$, $p = .01$, and accuracy, $t(19) = 2.72$, $p = .01$, than items followed by the classification cue. The difference in d' arose from a difference in hit rates: Classified stimuli had lower hit rates than stimuli for which participants indicated preference, $t(19) = 4.11$, $p < .001$, $d = 1.11$. There was no hint of a difference in false alarms, $t(19) < 1$. Typicality significantly correlated with hit rates in both study contexts—classification: $r(78) = .39$, $p < .0005$; preference: $r(78) = .28$, $p < .025$—indicating better memory for atypical items. When typicality was entered as a covariate in a general linear model to predict hit rates, the analysis revealed a marginally significant typicality by study condition interaction, $F(1, 152) = 3.19$, $p < .08$ (four outliers with standardized residuals greater than 2.4 *SD*s, less than 3% of all items, were removed from the analysis), indicating that classification had a larger effect on the most typical items.

Discussion

In Experiment 3, participants did not know what type of response they would be asked to make while viewing the pictures in the study phase. Nevertheless, it may be possible for preference judgments to produce deeper or more distinctive encoding by further processing the iconic memory store. This possibility was tested in Experiment 4 by including a pattern mask immediately after stimulus offset. Backward masking slowed responses in the study phase, suggesting that it was somewhat effective at terminating perceptual processing. Items followed by a classification cue were still remembered more poorly than items followed by a preference cue, suggesting that the difference between conditions did not arise from different degrees of encoding of iconic memory.

Experiment 5

In Experiments 3 and 4, poorer recognition memory was observed for labeled items even when participants did not know at the time of viewing whether they would be asked to label the item. Nevertheless, an alternative to the claim that this effect was due to a top-down effect of labeling remains. In all experiments presented so far, decision times were longer for preference judgments than for classification judgments. Decision times were also longer for atypical items than for typical items. If longer decision latencies for the preference judgments reflect more elaborate or distinctive encoding through rehearsal (Craik & Watkins, 1973; Woodward, Bjork, & Jongewar, 1973) or if greater effort produces higher arousal, which in turn produces better memory (Bradley, Greenwald, Petry, & Lang, 1992; cf. Hirshman, Trembath, & Mulligan, 1994), then the observed pattern of results can result from differences in encoding rather than from a postencoding distortion of the memory trace. The present results may therefore be explained by an account based on depth-of-processing mechanisms rather than by the proposed representational shift account.

Table 2
Mean Proportion (and Standard Deviation) of Hits and False Alarms, Mean Accuracy (Hits Minus False Alarms), and d′ for Experiments 4–5

Experiment and condition	Hits	False alarms	Accuracy	d'	Effect size of d' difference between conditions (Cohen's *d*)
Experiment 4	*		*	*	
Preference	.71 (.09)	.32 (.14)	.40	1.11	0.76
Classification	.62 (.14)	.40 (.19)	.23	0.64	
Experiment 5	*		*	*	
Preference	.83 (.10)	.41 (.16)	.43	1.29	0.59
Typicality	.75 (.13)	.40 (.16)	.35	1.00	

* $p = < .05$.

Two predictions made by the depth-of-processing account are contrasted with the representational shift account and tested in Experiment 5. First, if worse memory for overtly classified items in Experiments 1 through 4 resulted from less time spent encoding the items (as indicated by significantly shorter RTs in the classification condition), then according to a depth-of-processing account, category-related responses that took longer than preference judgments should yield better recognition memory. The representational shift account predicts that a category-related response would still yield poorer recognition memory despite a reversal in the relative RTs of the study conditions.

Second, the depth-of-processing account suggests that the correlation between hit rates and typicality arose from deeper or more distinctive encoding of atypical items (as suggested by longer classification RTs) rather than from top-down effects of the category labels (the attention-elaboration hypothesis; e.g., Erdfelder & Bredenkamp, 1998). A depth-of-processing account would therefore predict that if the correlation between the amount of elaboration (using the proxy of study decision RTs) and typicality is disrupted, the correlation between typicality and memory will also be disrupted (see Craik & Tulving's [1975] Experiment 5 for similar reasoning). In contrast, the representational shift account predicts that a greater degree of category-related processing would *decrease* memory insofar as it introduces a greater degree of distortion of the visual features by category-level information. Because classification RTs correlate with typicality, using classification as a study condition cannot discriminate the two accounts. Classification response RTs are shortest for the most typical items, so it is not clear whether poor memory following their classification is due to representational shift or poor encoding as suggested by shorter RTs (though the correlation between typicality and hit rate for the classified items remains when study RTs are partialed out).

In Experiment 5, classification judgments were replaced by typicality judgments. Analyzing RTs of previously collected typicality ratings revealed an inverted U curve, with intermediate typicality ratings having the longest RTs. Thus, the depth-of-processing account would predict that intermediate typicality judgments should result in the strongest or most distinctive encoding and thus better memory. In contrast, the representational shift account predicts poor memory (lower hit rates) for these items despite an arguably greater level of encoding. It should be noted that although RTs are not always correlated with depth of processing (Craik & Tulving, 1975), greater study RTs in the preference condition remain the only a priori reason to suspect that preference judgments produced deeper encoding. If the differences in recognition memory are merely a consequence of study times, then switching the direction of the difference in study RTs should also reverse the direction of recognition memory.

Method

Eighteen participants (mean age = 23.8 years, *SD* = 3.3 years) gave informed consent and received $7 for participation. Two participants were eliminated for failing to follow instructions (one used the wrong buttons in the test phase, another used only the extreme typicality responses). The study and test items were chairs and lamps, identical to those in Experiment 1. The procedure was similar to that of Experiment 1 with the following exceptions: Instead of making classification judgments, participants were asked to rate stimuli on a scale of 1–5, with 1 corresponding to a very typical item and 5 corresponding to a very atypical item. Half of the stimuli were presented in this typicality condition, and the other half were presented in a preference context, as in Experiment 1. The order of presentation (chair first, lamp first, preference first, typicality first) was fully counterbalanced between participants. In addition to being blocked by condition, stimuli were blocked by category to avoid participants having to first judge whether the object was a chair or lamp, in which case the resulting RTs would reflect a combination of categorization and typicality-judging processes.

Each block was preceded by an instruction telling participants whether they should be making preference or typicality responses and, if the latter, what category they would be seeing. For instance, before the *chair–typicality* block, participants were instructed that they would now be making typicality judgments for chairs. Following the offset of each stimulus, a response cue appeared showing participants the 1–5 typicality scale and the prompt "How typical was that chair?" Participants were instructed to use the 1–5 number keys on the keyboard for responding. For preference blocks, participants were prompted with the question "Do you like this object?" along with a reminder of which buttons to use for the preference judgment (the green button on the gamepad for *like* and the red button for *don't like*—the same buttons used for preference judgments in the preceding experiments). The response window was lengthened to 1,800 ms to accommodate the longer RTs produced in the typicality judgments while at the same time requiring participants to respond relatively quickly. Finally, as in Experiment 4, participants were not told to try to remember the stimuli or that a test would follow the study phase.

Notice that in the preference condition participants had two alternative responses, *like* and *don't like*, whereas in the typicality condition participants were using a 5-point Likert scale. The latter is designed to be a more difficult task with longer RTs and involving arguably deeper encoding. According to a depth-of-processing account, these items should be best remembered. According to the representational shift account, these items should be remembered more poorly, even though the study context in which they were encountered involved more taxing judgments. Thus, any confound introduced by a difference in response options would favor the depth-of-processing account over the proposed representational shift account.

Results

Study Phase

RTs for typicality judgments were significantly longer than RTs for preference judgments ($M_{\text{preference}}$ = 753 ms, *SD* = 121 ms; $M_{\text{typicality}}$ = 978 ms, *SD* = 110 ms), pairwise *t* test, $t(15) = 6.41$, $p < .0005$. The typicality RTs were characterized by an inverted U curve, with the intermediate typicality ratings producing the longest RTs (Figure 4, top). A polynomial regression analysis predicting RT from typicality response revealed a significant contribution from the quadratic component ($F = 12.29$, $p = .001$, $R^2 = .28$), further confirming that the RT curve was nonlinear.

Test Phase

Recognition memory for items studied in the typicality blocks was significantly lower than for items studied in the preference blocks, as revealed by a comparison of d', $t(15) = 2.61$, $p < .025$, and accuracy, $t(15) = 2.16$, $p < .05$. As before, the difference arose from a difference in hits, $t(15) = 3.80$, $p < .005$, $d = 0.74$. There was no significant difference in false alarms, $t(15) < 1$.

In the typicality condition, hit rates were lower for items given the intermediate responses, 2–4, than for those given 1 and 5 ratings, $F(1, 298) = 4.49$, $p < .05$ (Figure 4, middle). To compare the relationship between typicality and hits in the two conditions, a regression analysis was performed with hit rates in the typicality and preference conditions as the outcome variable and average item typicality as provided by the current participants as the predictor. The results are shown in Figure 4 (bottom). Hit rates following typicality judgments were best predicted by a quadratic function, $F(2, 37) = 3.24$, $p < .05$, $R^2 = .15$. Both the linear and the quadratic components were significant ($\beta_{\text{typicality}} = -.54$, $p < .05$; $\beta_{\text{typicality}}{}^2 = .10$, $p < .025$). Typicality did not significantly predict hit rates following preference judgments, $F(2, 37) = 1.03$, $p = .37$, $R^2 = .05$. An additional analysis of items with intermediate versus extreme typicality values can further clarify this analysis. For items within the middle range of typicality (second and third quartiles), there was a significant difference between the preference and typicality conditions, $t(19) = 3.15$, $p < .01$. For the most typical and atypical items (first and fourth quartiles), there was no significant difference in hit rates between the study conditions, $t(19) < 1$. In all cases, study RTs for typicality judgments were much longer than for preference judgments (Figure 4, top), and so an account based on depth of processing would predict that if any difference in memory between the conditions was found, it should be in the direction of superior memory in the typicality condition, the opposite of the present finding.

Discussion

Making judgments related to the item categories (typicality ratings) resulted in poorer memory compared with preference judgments even though typicality judgments were a more difficult task, arguably requiring deeper processing of the stimuli. This finding suggests that the memory differences observed in Experiments 1 through 4 were not due to differences in study times.

The failure to find an effect of study condition for the atypical items is consistent with the findings from the earlier experiments. Why was memory for the most typical items affected by classification judgments in Experiments 1 through 4 but not by typicality judgments in the present experiment? One explanation is that the influence of the category label may have been minimized in the present experiment both by not requiring an overt classification response and by blocking the stimuli by category. This would produce a smaller difference between study conditions (as confirmed by the smaller effect size in the current experiment compared with Experiments 1 through 4; cf. Tables 1 and 2). This decrement would be particularly evident for the most typical items, for which the effect of overt classification is predicted to be the strongest by the representational shift account. In addition, the potentially greater category priming induced by the blocked design of the present experiment may have altered the interaction between the category label and item representations in ways not fully captured by the present account.

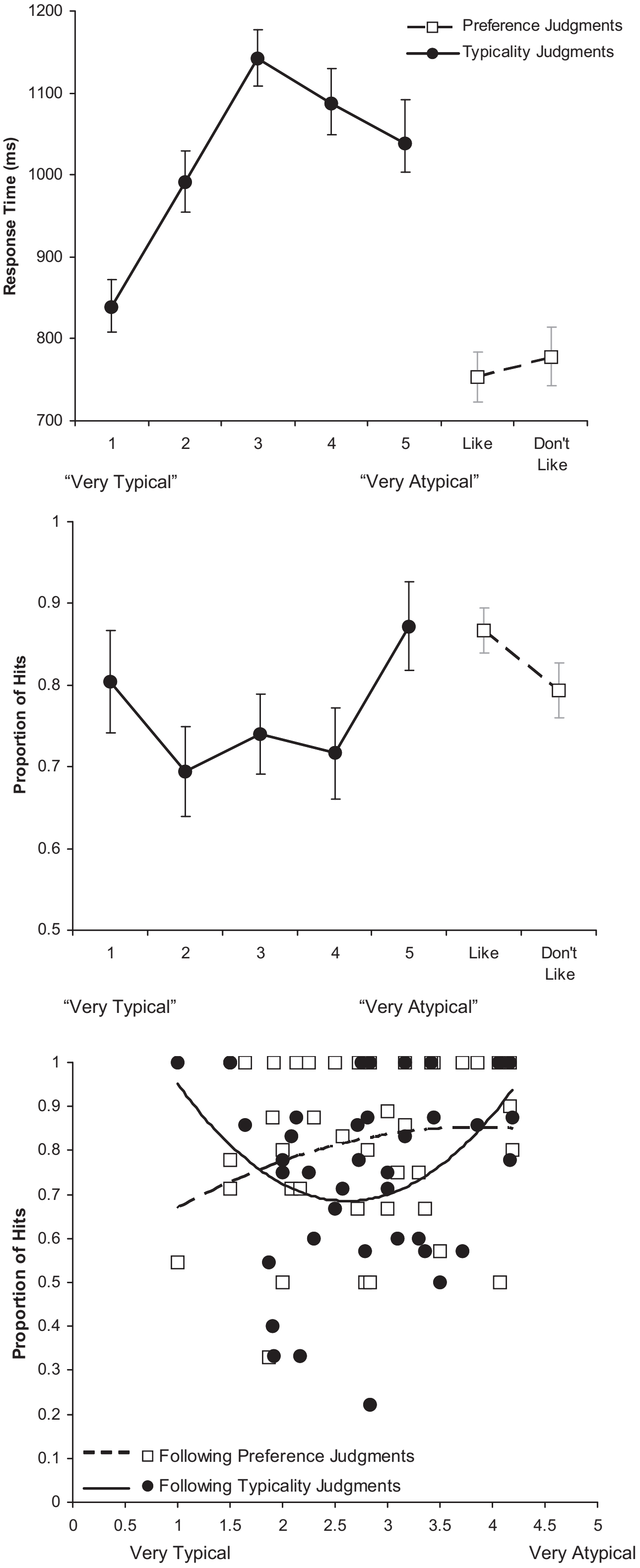

Figure 4. Top: Experiment 5 study phase reaction times as a function of response type. Middle: Subsequent hit rates for the items as a function of response type. Bottom: Proportion of hits by stimuli and condition. Points represent individual stimuli. Closed symbols represent items studied in the typicality context, and open symbols represent items studied in the preference context. Stimuli of intermediate typicality have the lowest hit rates following typicality judgments even though intermediate typicality judgments have greater reaction times than more extreme judgments.

Experiment 6

According to the representational shift account, lower hit rates for items classified with category labels arise from a representational mismatch caused by top-down effects of the label. Category labels are predicted to distort the items' representation by augmenting the representation of the exemplar with category-typical information. One hypothesized consequence of such a distortion is that items affected by the category label should be judged as being more typical. Crucially, the degree to which an item is rated as being more typical in the context of a label should vary as a function of its original typicality. Because the exemplar-to-label association is strongest for the typical items, the predicted effect of the labels is also strongest for the typical items, leading to the somewhat counterintuitive prediction that category labels should make already typical items even more typical. Conversely, ratings of atypical items should be minimally affected by category labels. To gain an intuition for why this should be the case, consider a chair without armrests that, through the influence of the chair category, comes to be represented, mistakenly, as having armrests. Because armrests are a feature highly diagnostic of the chair category, a chair with armrests is likely to be rated as more typical than a chair without armrests. A comparison of typicality ratings of chairs with and without armrests supports the intuition that chairs with armrests are more typical than chairs without armrests: Mean typicality ratings of chairs with and without armrests were, respectively, 1.24 and 3.06, $F(1, 18) = 6.89$, $p = .02$.

Experiment 6 was designed to test the idea that the hypothesized representational shift produced by category labels can be observed as an increase in perceived typicality. It was reasoned that merely seeing the category label ("chair" or "lamp") while making a typicality rating would activate features typical of the labeled category more strongly than a typicality response in the absence of seeing the name of the object category. Moreover, if the effects of category labels are mediated by typicality, then the difference in ratings should be greatest for the most typical items because it is they that are most strongly associated with the category label and so are most affected by it.

Method

Twenty-one participants (mean age = 19.4 years, *SD* = 1.5 years) gave informed consent and received course credit for participation. Two participants were eliminated for giving extreme responses leading to mean ratings more than 2.5 *SD*s below the mean. One participant was eliminated for having unusually fast decision times (2.5 *SD*s below the mean), which suggested a greater degree of anticipatory planning than other participants. The study items were chairs and lamps, identical to those used in Experiment 5. The procedure was similar to that of Experiment 5 with the following exceptions: There was no test phase. Participants rated the typicality of all rather than half of the studied items. They used a 5-point Likert scale, identical to that used in Experiment 5. Following the brief (300-ms) presentation of each item, the 1–5 rating scale appeared together with one of two prompts. In the *labeled* condition, the prompt labeled the category of the object just shown: "Please rate the typicality of the [chair/lamp] you just saw." In the *generic* condition, the prompt was "Please rate the typicality of the object you just saw." Unlike Experiment 5, the items were not blocked by category. Chairs and lamps were intermixed and randomly assigned to the *labeled* and *generic* conditions, such that while viewing each object participants did not know ahead of time the exact prompt that would follow. For paradigm consistency, each object was presented a total of two times, as in the previous studies. The pairing of items to conditions was counterbalanced, so that for each participant who rated a given half of the items in the labeled context, there existed a matched participant who rated the same half of the items in the generic context. As a result, each item was rated by an equal number of participants in both labeled and generic contexts.

Results

Typicality ratings in the labeled and generic conditions were highly correlated $r(38) = .95$, $p < .0005$. To analyze differences in responses between the labeled and generic conditions a repeated measures analysis of variance was performed with items as a random factor, and condition and block (first vs. second exposure) as fixed factors. Mean typicality ratings for the two conditions are shown in Figure 5. The analysis revealed a main effect of block, $F(1, 39) = 13.56$, $p < .001$, with participants rating objects as more typical on the second exposure compared to the first. The main effect of condition did not reach significance, $F(1, 39) = 2.08$, $p = .16$. There was, however, a significant block by condition interaction, $F(1, 39) = 4.78$, $p < .05$. Post hoc tests showed that in the first block, items were rated as being more typical when presented in the labeled condition ($M_{\text{generic}} = 3.08$, $SD = 0.93$; $M_{\text{labeled}} = 2.93$, $SD = 1.08$), pairwise *t* test, $t(38) = 2.09$, $p < .05$. The differences between the conditions disappeared by the second block ($M_{\text{generic}} = 2.87$; $M_{\text{labeled}} = 2.85$), $t(38) < 1$.

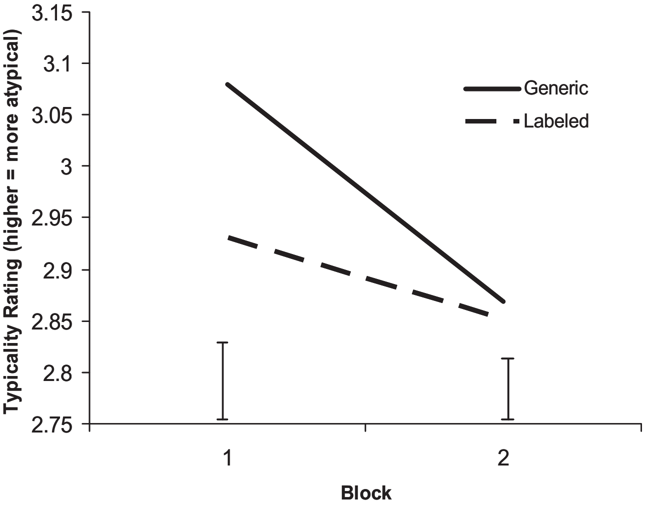

Figure 5. Mean typicality ratings for items in the generic and labeled conditions for the first and second presentations (lower values = more typical). Error bars show one standard error of the within-item difference score (generic minus labeled).

To determine whether the already typical items underwent the greatest increase in typicality, a difference score was computed by subtracting the mean typicality ratings for each item in the labeled condition from the same item's rating when it was presented in the generic condition. Both presentation blocks were included in the analysis. This difference score was correlated separately with two measures of typicality: the original typicality ratings from a separate group in which items were blocked by categories (used in Experiments 1 through 5), typ_{original}, and for reliability, a second measure, typ_{average}, computed by averaging the typicality measurements for each item in the labeled and generic conditions of the present study. The correlation between the typicality difference score and typ_{original} was significantly negative, $r(38) = -.32$, $p < .05$. The correlation between the difference score and typ_{average} was also negative, $r(38) = -.35$, $p < .05$. The negative correlation coefficients indicate that the most typical items underwent the largest increase in typicality in the context of the category label (the labeled condition).

An analysis of response times with stimuli as a random factor and presentation block and condition as fixed factors revealed a significant effect of block, $F(1, 39) = 71.41$, $p < .0005$. Unsurprisingly, RTs were faster on the second block. There were no effects of condition, $F(1, 39) < 1$ ($M_{\text{generic}} = 1{,}143$ ms, $SD = 115$ ms; $M_{\text{labeled}} = 1{,}126$ ms, $SD = 145$ ms). The block by condition interaction was not significant, $F(1, 39) < 1$.

Discussion

Though participants were most likely automatically categorizing the images as they appeared, and though the wording of the typicality prompt was ancillary to the response, the inclusion of the category name ("chair" or "lamp") in the response prompt affected the typicality ratings; items followed by a category label were judged to be more typical. Rather than a simple shift in the mean responses, as might be expected if adding the label changed the response bias, the magnitude of change in typicality responses was itself mediated by item typicality. Labels produced the greatest increase in typicality for already typical items.

There are several possible explanations for this finding. Participants may have been confused about the category of some of the items, and this confusion was corrected by presenting the category label in the subsequent prompt. As can be seen in Figure 1, images of chairs and lamps shared few common features and were thus hard to confuse. To be sure, some items were quite atypical. If seeing the category name reminded the participants to which category the object belonged, then the difference in the typicality ratings between the labeled and generic conditions should have been largest for the atypical items. However, the opposite pattern was observed: Responses for the typical, but not atypical, items were affected by the presence of the label.

The finding that typical items were more affected by category labels is accommodated by the representational shift account, according to which the top-down feedback produced by a category label affects most strongly items possessing attributes that are associated with the object's category (see, e.g., Discussion of Experiment 2). The label effectively augments the values of these attributes with more typical values. Thus, an already typical item is perceived to be an even more typical member of the category. The finding of overall greater judged typicality on the second presentation of items supports the general idea that participants' typicality ratings are based not only on the immediate presentation of an item but on its constructed representation, which is apparently affected by familiarity (first vs. second presentation), as well as the presence of a category label. It should be noted that this finding of increased typicality for the items judged as being the most typical indicates that these items are in fact not *prototypical* in the true sense of the word. An argument can be made that a true prototype of a chair (or any complex object) is never encountered because such an object would have to possess all of the most typical features simultaneously.

The finding that the effect of labels on representations is strongest for the most typical items is in line with the findings from Experiments 1 through 4 but at odds with Experiment 5, in which labeling produced the largest representational shift for items of intermediate typicality (insofar as a lower hit rate is a proxy measure of representational shift). A possible cause of this discrepancy is that blocking stimuli by category in Experiment 5 resulted in some type of category priming, reducing the effect of the label for the most typical stimuli (the stimuli containing the most typical, and so most strongly primed, features).

General Discussion

Experiments 1 through 5 showed that overtly classifying familiar objects (chairs, tables, and lamps) with their basic-level names resulted in worse recognition performance than when the items were not overtly classified. In accordance with the idea that classification results in a study-to-test mismatch, overtly classified items had lower hit rates than items not overtly classified. This decrease in hit rate was strongly mediated by the typicality of the labeled items, with the largest effect for the most typical objects. An additional prediction of this account was that recognition of the most unambiguous items (e.g., the most untablelike chairs) should be more impaired by classification than recognition of ambiguous items. This prediction was confirmed in Experiment 2.

Experiments 3 and 4 tested the possibility that the observed effects might be due to strategic differences in visual or attentional processing. Because participants in Experiments 1 and 2 knew what kind of response they would be making for each item, it was possible that they focused on different visual features, perhaps paying more attention to perceptual details for preference judgments and structural details for classification judgments. If this were the case, the observed differences in memory might be due to classification resulting in a different encoding strategy rather than producing a representational shift. Experiment 3 tested this possibility by randomly varying the response cue after each item. Experiment 4 added a mask following stimulus offset to limit the degree to which participants could rely on iconic memory in encoding the stimulus. Even though participants now did not know what type of response would be required, recognition was still worse for the classified items regardless of whether the stimulus was masked. In Experiments 1–3 memory for the classified items was poorest despite participants being instructed to try to remember as much as possible about each item. Experiment 4 obtained the effect while omitting this instruction.

Experiment 5 sought to further discriminate the representational shift account of the present results from an explanation based on a depth-of-processing account. Experiment 5 showed that a study context requiring a type of category-related response (typicality judgments) resulted in poorer memory than a response unrelated to the category (preference judgments) despite the former being more effortful. This contradicts the depth-of-processing prediction of better memory following arguably more elaborate encoding while supporting the prediction of the representational shift account of poorer memory following greater engagement with the item's category.

Correlating item typicality and ambiguity with subsequent recognition performance revealed that labeling had the greatest effect on the most typical and unambiguous items, that is, the items most strongly associated with the category label (though see Experiment 5). Thus, memory for the items that were easiest to label (the typical and unambiguous stimuli) was most affected by labeling. A possible mechanism of this finding is described in the following section.

The overt classification condition was contrasted here with preference judgments. The latter served as a proxy for a *no-label* condition by requiring participants to make a speeded response unrelated to the item's category and thus, arguably, partially suppressing the top-down influence of the category on the encoded representation. To establish with more certainty that the effect is truly about labeling or overt classification, comparisons need to be made between memory following labeling and memory following numerous other types of judgments unrelated to the category, or alternative means suppress the effects of the category label. Because the present experiments used preference judgments as a control condition, caution should be exercised in generalizing the results to other contexts.

The Representational Shift Account of Labeling Effects on Memory

The representation shift account attempts to explain why labling familiar items results in poorer within-category recognition memory, particularly for the most typical and unambiguous items. This account requires treating items not as atoms but as collections of features in a high-dimensional space. Some dimensions (and features within those dimensions) are reliably associated with a particular category: for instance, armrests with chairs but not tables. Others are more weakly associated: for example, having a back with chairs (not all chairs have backs). Still other dimensions are not predictive of a particular category because they are either irrelevant to the category (e.g., color for chairs and tables) or common to many categories and so do not help to distinguish one category from another in a given context (e.g., both chairs and tables tend to have flat surfaces). A category name like "chair" does not capture idiosyncratic properties of any particular chair. In the course of experience with labeling various chairs with "chair," the label becomes most strongly associated with features most commonly associated with chairs—the typical properties—and becomes dissociated from features not reliably associated with the label—the atypical properties. For instance, color does not predict whether an item is a chair or a table, but the presence of armrests not only makes it likely that an object is a chair and not a table but also predicts other features, such as the presence of a back.

When a category name is activated following a presentation of an object, it is hypothesized to augment through top-down feedback the features activated by the bottom-up input from the recently experienced item. The resulting representation of a labeled item thus combines the idiosyncratic features of a particular item, with features typical of the labeled category. This distortion produces a study-to-test mismatch that makes old items presented at test seem newer, yielding a reduced hit rate. Greater augmentation of the visual (feature) representation is associated with a greater mismatch and hence lower hits. The representational shift is greatest for typical and unambiguous items for two reasons. First, these items produce the most reliable bottom-up activity to the category label (e.g., as indicated by faster labeling of typical/unambiguous items). Weakly activated labels produce less top-down modulation than strongly activated labels. But this is not the whole story. Typical items, by definition, have a greater proportion of category-relevant features that are strongly linked to the label. However, these features do not necessarily have the most typical *values*. For instance, imagine an unambiguous chair with armrests (typical feature) that are of an unusual shape (atypical value). The top-down feedback from the category label may then distort the item's representation to essentially confabulate an item with typical armrests. Compare this scenario with that of studying an atypical or ambiguous item. Such an item has fewer category-typical features, and so even if the label is allowed to activate fully, there is little for its feedback to modify because the item's original representation has few features that are strongly associated with the label. Consequently, there is less top-down feedback and the item is remembered with greater fidelity.

If labels make objects more typical, then participants may judge objects as being more typical in the context of the label. This is precisely what was observed in Experiment 6. Rather than a simple shift in criterion, judging typicality in the context of labels made already typical items even more typical. Judgments of atypical items were unaffected by labels, arguably because they possessed fewer category-typical attributes that could be modified by the label.

It is important to note that the prediction of poorer memory following overt classification holds only for *within-category* recognition. Indeed, the present account would predict that labeling study items might produce superior memory in a between-category task, because one effect of the label feedback is to "clean up" the studied items to make their representations less noisy and more categorical. Consistent with this general idea, we have found that learning labels for novel categories facilitates the learning of the categories compared with a condition in which participants have equivalent experience with supervised categorization of the items but without the benefit of labels (Lupyan, Rakison, & McClelland, 2007). The results are consistent with the notion that activation of labels results in more robust category attractors. Though beneficial in a category-learning task (where within-category differences need to be abstracted), it would be detrimental in a task requiring faithful representations of individual exemplars, as was necessary in the present studies.

Representational Shift Account Versus Depth of Processing

The representational shift account argues for an augmentation of study items with higher level category information. The initial encoding of the item is likely to depend on the desired response, at least when the response is known ahead of time (Experiments 1 and 2), but critically, the effect found here is argued to not depend solely on differences in initial encoding. The postencoding nature of the proposed mechanism is at odds with encoding-level accounts such as depth of processing. In this section it is argued that the ability of depth of processing to account for the present results is suspect on both theoretical and empirical grounds.

According to depth-of-processing accounts, semantic processing, necessary for labeling an object, produces stronger memory traces (Craik & Lockhart, 1972). Subsequent clarifications of this framework predicted that different study tasks result in encoding that may be of comparable strength but vary in distinctiveness, with more distinctive items being easier to access in a subsequent recognition task (Lockhart et al., 1976). Although the presence of depth-of-processing effects in memory tasks is incontrovertible, it is unclear how an account based purely on differences in encoding can account for the present results. The first problem with a depth-of-processing-based account is that it is not clear which condition would be predicted to produce the more distinctive or more elaborate encoding. On the one hand, labeling a chair as a "chair" may involve comparing the given chair with other chairs in memory and focusing on their differences, and so would be predicted to result in a more distinctive memory trace. Alternatively, it is the encoding of perceptual details produced by a perceptual judgment such as preference that may produce a memory advantage (Marks, 1991). Alternatively still, if preference judgments are accompanied by idiosyncratic justifications of why an item is liked or not, it is the preference condition that may result in the more distinctive encoding (possibly producing a self-reference effect). It seems doubtful, however, that speeded judgments asking participants whether they like an object require the same level of introspection as asking whether a particular attribute describes it—the kind of manipulation that produces a memory advantage through a self-reference effect (Symons & Johnson, 1997).

The trouble with depth-of-processing accounts is that they do not allow a way to independently assess degree of encoding except with reference to the testing performance (though see Kapur et al., 1994, for a neuroimaging measure of encoding depth). If recognition memory is theorized to be purely a function of encoding strength and specificity, then predicting encoding strength from memory is unproblematic. If, however, recognition memory is the result of a representational match between match study and test items, then recognition performance becomes a function of both encoding effects and any distortion of the encoded representation that may occur from study to test. The appropriateness of the test session with respect to the study condition is also relevant (Morris, Bransford, & Franks, 1977).

If study RTs are used as a proxy for the degree of encoding, the finding of better recognition memory for preference (longer study RTs) and atypical items (longer classification RTs) can be argued to arise purely from differences in encoding. Experiment 5, however, shows that even when category-related responses produce more elaborate encoding as suggested by longer RTs, memory is still poorer for items for which a category-related response was provided, contrary to accounts predicting simple differences in initial encoding. If study RTs are rejected as a proxy for the degree of encoding, then there is no a priori reason to think that classification produces less elaborate or less distinctive encoding than preference judgments, and depth-of-processing accounts again cannot predict the present results.

Shift-to-Prototype and Perceptual Magnet Effects

Depth-of-processing accounts do not make clear predictions regarding how typicality and ambiguity should relate to effects of labels on memory. Two theories, however, do make predictions with regard to typicality that are relevant to the present findings. These are the shift-to-prototype effect (Huttenlocher, Hedges, & Duncan, 1991) and the perceptual magnet effect (Kuhl, 1994). Both would predict that if labels are most strongly associated with category prototypes, then labeling items may act to shift their representations closer to the prototype. Specifically, Huttenlocher et al. (1991) predicted that items closest to the prototype undergo a smaller representational shift (reporting bias) than the less typical items. Thus, the greater effect of classification should be for the items farther from the prototype, that is, the atypical items rather than the typical items. The present results reveal the opposite pattern: It is the more typical items that are most affected by classification and that, by the logic of the representational shift account, undergo the largest shift.

This discrepancy between the present results and the predictions of the shift-to-prototype theory and the perceptual magnet effect can be explained by noting that in these accounts, the prediction of whether labeling should affect representations of typical or atypical items depends entirely on the nature and shape of the attractors (or "magnets"). Using the magnet metaphor, we can ask how a magnet's (category prototype—activated by the label) effect on a thumbtack (studied item) varies as a function of the distance between the two. Greater movement would correspond to a larger representational shift and so poorer recall. A thumbtack close to the magnet would be under a stronger influence of the magnet, but because it is already close to the magnet, it would not move very much (Scenario 1). A thumbtack slightly farther from the magnet but nevertheless under its influence might move a greater distance (Scenario 2). But a thumbtack farther still would be outside the pull of the magnet and so would not move at all (Scenario 3). The prediction from perceptual magnet and shift-to-prototype effects depends on which pair of scenarios corresponds to the current case. If Scenario 1 corresponds to typical items and Scenario 2 to atypical, the prediction is poorer memory for labeled atypical items and better memory for labeled typical items. If Scenario 2 corresponds to the typical items and Scenario 3 corresponds to the atypical items, the prediction is reversed. The predictions therefore depend on defining the full feature space of the items. Whereas prototypes are easily definable in low-dimensional spaces such as colors or oriented lines, it is altogether unclear what a true prototype of a real-world object like a chair looks like. There is no guarantee that items rated as being the most typical in any set of objects are truly at the center of the category, because no item is likely defined on all possible dimensions, leaving room for the label to further augment its representation. The results of Experiment 6 further elaborate this point: Rating typicality in the context of a label resulted in a further increase in typicality for the already typical items, thus demonstrating that items rated as the most typical are in fact not at the category "center" and indeed subject to further shifting.

Why No Differences in False Alarms?

One conception of labels is that they are features of objects (Postman, 1955; Sloutsky & Lo, 1999; the SINC model of Sloutsky & Fisher, 2004a; cf. Gibson & Gibson, 1955). According to this view, objects that share a label are made more similar and possibly more confusable (Robinson, 1955; e.g., Katz, 1963). This account would predict lower recognition rates for the classified (i.e., labeled) items, with the difference expressed in false alarms (Sloutsky & Fisher, 2004b; Sloutsky, Lo, & Fisher, 2001). In contrast, the present studies reveled a difference in hits. Why did labeling not result in greater confusability of items (i.e., higher false alarms in the classification condition)? There are several possibilities. In Experiments 1–3 participants were explicitly told to pay careful attention to, and remember, each item—instructions that likely discouraged them from merely treating each item as a member of a category. The presence of only two categories meant that paying attention to just the category was trivial—participants clearly had no trouble remembering which categories they had seen during the study phase. Even when no memory instructions were given, participants had an incentive to encode more than just the item's category, because they had to provide preference judgments for some of the items. In contrast, tasks such as Sloutsky and Fisher's (2004b) induction condition promote little incentive to pay attention to the individual features of objects and thus produce the expected result of category-level encoding (though seemingly not for children), in which case a difference in false alarms is expected and observed. In short, encoding-level accounts predict greater false alarms for classified objects insofar as classification leads to a "coarser" encoding of the study item. There is no evidence that participants in the present experiments treated study items in this kind of categorical way, and so the lack of reliable differences in false alarms is not surprising.

Representational Shift Due to Labels as a Mechanism for the Verbal Overshadowing Effect

The representational shift account proposed to explain the current results may be relevant to explaining the verbal overshadowing effect reported by Schooler and Engstler-Schooler (1990). These authors found that memory for faces is decreased if people are asked to verbally describe the studied face before attempting to choose it among alternative faces. The effect has been extended to other objects (e.g., cars; Brown & Lloyd-Jones, 2003) and to other modalities (e.g., wine tasting; Melcher & Schooler, 1996). As in the present experiments, the effect of verbal description is that of a reduction in hits. Schooler (2002) has proposed that verbalization produces a "'transfer inappropriate processing shift' whereby the cognitive operations engaged in during verbalization dampen the activation of brain regions associated with critical non-verbal operations" (p. 989). The current results propose an alternative explanation of the effect. It is possible that reduced recognition of faces after a verbal description results from a representational shift caused by words in the produced description referring to entire categories. For instance, describing someone as having "brown eyes" results in representing the eyes as a more prototypical brown rather than a particular shade of brown. A prediction arising from this account is that typical objects should be more subject to verbal overshadowing effects because they are more likely to possess features better captured by category labels. Consistent with this prediction, a recent study by Wickham and Swift (2006) found that typical faces, but not distinctive faces, are subject to verbal overshadowing and that the effect of overshadowing is attenuated by articulatory suppression.

An additional factor contributing to the verbal overshadowing effect of faces, cars, and wines is that such stimuli are uniquely defined not by single cues but rather by configural differences. All faces have noses, and all cars have wheels. It is the relationship between these features that defines a face or a car at an individual level. Most languages seem far better suited for describing features and actions rather than configurations (e.g., in English, nouns and verbs, but not prepositions, are open classes allowing introduction of new words). Hence, providing verbal descriptions for items defined by configurations may produce distorted representations, but providing verbal descriptions for items defined by unique features may actually help in recall. This prediction remains to be tested.

Further Implications

The effect of labels on recognition memory as revealed in the present studies is pronounced in magnitude and can be observed for highly familiar objects in a within-subject design. The present experiments required participants to recognize having seen exact members of a particular category. Classification using category labels was detrimental to this task. But while recognition of particular category members is undoubtedly important, as when one searches for one's own car rather than just a car, most categorization tasks are best performed at a higher, more general level that requires abstracting over idiosyncrasies of particular category members. This leads to a striking prediction: Deficits in naming ought to lead to deficits in categorization. Support for this prediction can be found in the literature on aphasia. As a group, aphasic patients, particularly those with word-finding difficulties, are impaired on sorting colors (Basso, Faglioni, & Spinnler, 1976; De Renzi, Faglioni, Scotti, & Spinnler, 1972) and sorting geometric figures or familiar objects according to common attributes such as color, shape, size, and function (De Renzi, Faglioni, Savoiardo, & Vignolo, 1966; Gainotti, Carlomagno, Craca, & Silveri, 1986; Hjelmquist, 1989; Kelter, Cohen, Engel, List, & Strohner, 1977), and they perform poorly on classical sorting tasks, such as the Weigl Sorting Task (De Renzi et al., 1966; De Renzi, Spinnler, Scotti, & Faglioni, 1972; Koemeda-Lutz, Cohen, & Meier, 1987) and the Wisconsin Card Sort (Baldo et al., 2005). Whereas poor categorization performance might be predicted when word-finding difficulties are accompanied by comprehension problems, many sorting and categorization impairments, particularly classification along perceptual dimensions such as size, shape, and color, exist even in patients who show relatively intact semantic categories (Basso et al., 1976; Cohen, Kelter, & Woll, 1980; Davidoff & Roberson, 2004; De Renzi et al., 1972; Hjelmquist, 1989; Kelter, Cohen, Engel, List, & Strohner, 1976; Roberson, Davidoff, & Braisby, 1999), and in one case, adding labels helped bring aphasic patients closer to normal performance (Koemeda-Lutz et al., 1987). This is further evidence that language plays an active role in normal functioning—a view compatible with the aphasiologist's Kurt Goldstein's dictum "Language is not only a means to communicate thinking; it is also a means to support it, to fixate it," and "defect in language may thus damage thinking" (Goldstein, 1948, p. 115).

What about the converse? Can deficits in naming produce better within-category recognition memory? To my knowledge, there have been no studies on within-category memory in aphasic patients. However, a recent study by Roberson (2006) found that whereas low-functioning autistic children are impaired in their ability to make category-appropriate color choices in response to a given color (expected if labels facilitate category-level decisions), they actually exceed the performance of normally developing children in a subsequent unexpected within-category memory test. It seems that the use of category labels by normally developing (and high-functioning autistic) children leads them to fail in accurately encoding or maintaining a representation of within-category color distinctions.

Conclusion

Counter to the common conception of words as merely a means of communicating already formed ideas (Fodor, 1975; Li & Gleitman, 2002; Pinker, 1994), the present experiments demonstrated that simple labeling of familiar objects results in lower subsequent recognition: Participants fail to recognize previously seen labeled items compared with items not overtly labeled. Although participants likely implicitly classified all of the studied items, the overt labeling response is hypothesized to increase the influence of the category compared with a condition in which a response unrelated to the category is required.

Control experiments showed that although accounts based on encoding differences (e.g., depth of processing) may have contributed to the present findings, such accounts cannot fully explain the present findings. The findings are explained through a representational shift account in which labeling produces a representational mismatch between the item as encoded at study and the original item presented at test. Naming a familiar item is hypothesized to engage top-down feedback, augmenting the representation constructed through bottom-up processing with top-down information—a kind of conceptual filling-in effect (Bransford & Franks, 1971; Franks & Bransford, 1971).

Memory for typical and unambiguous items is most compromised by labeling because labeling these items results in the strongest top-down feedback, both because they activate the label most strongly and because typical/unambiguous items have a greater proportion of features that can be modified by the top-down feedback—that is, those strongly associated with the category label.

References

Astley, S. L., & Wasserman, E. A. (1992). Categorical discrimination and generalization in pigeons: All negative stimuli are not created equal. *Journal of Experimental Psychology: Animal Behavior Processes, 18,* 193–207.

Astley, S. L., & Wasserman, E. A. (1998). Novelty and functional equivalence in superordinate categorization by pigeons. *Animal Learning & Behavior, 26,* 125–138.

Bachmann, T., & Allik, J. (1976). Integration and interruption in masking of form by form. *Perception, 5,* 79–97.

Balaban, M. T., & Waxman, S. R. (1997). Do words facilitate object categorization in 9-month-old infants? *Journal of Experimental Child Psychology, 64,* 3–26.

Baldo, J. V., Dronkers, N. F., Wilkins, D., Ludy, C., Raskin, P., & Kim, J. Y. (2005). Is problem solving dependent on language? *Brain and Language, 92,* 240–250.

Basso, A., Faglioni, P., & Spinnler, H. (1976). Nonverbal color impairment of aphasics. *Neuropsychologia, 14,* 183–193.

Bloom, P. (2001). Controversies in the study of word learning: Response. *Behavioral and Brain Sciences, 24,* 1124–1134.

Bloom, P., & Keil, F. C. (2001). Thinking through language. *Mind & Language, 16,* 351–367.

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, and Cognition, 18,* 379–390.

Bransford, J. D., & Franks, J. J. (1971). Abstraction of linguistic ideas. *Cognitive Psychology, 2,* 331–350.

Brown, C., & Lloyd-Jones, T. J. (2003). Verbal overshadowing of multiple face and car recognition: Effects of within- versus across-category verbal descriptions. *Applied Cognitive Psychology, 17,* 183–201.

Carmichael, L. C., Hogan, H. P., & Walters, A. A. (1932). An experimental study of the effect of language on the reproduction of visually perceived form. *Journal of Experimental Psychology, 15,* 73–86.

Carruthers, P. (2002). The cognitive functions of language. *Behavioral and Brain Sciences, 25,* 657–674.

Clark, A., & Karmiloff-Smith, A. (1993). The cognizer's innards: A psychological and philosophical perspective on the development of thought. *Mind & Language, 8,* 487–519.

Cohen, R., Kelter, S., & Woll, G. (1980). Analytical competence and language impairment in aphasia. *Brain and Language, 10,* 331–347.

Coltheart, M. (1983). Iconic memory. *Philosophical Transactions of the Royal Society of London, Series B, 302,* 283–294.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: Framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671–684.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and retention of words in episodic memory. *Journal of Experimental Psychology: General, 104,* 268–294.

Craik, F. I. M., & Watkins, M. J. (1973). Role of rehearsal in short-term memory. *Journal of Verbal Learning and Verbal Behavior, 12,* 599–607.

Davidoff, J., & Roberson, D. (2004). Preserved thematic and impaired taxonomic categorisation: A case study. *Language and Cognitive Processes, 19,* 137–174.

Dennett, D. C. (1994). The role of language in intelligence. In J. Khalfa (Ed.), *What is intelligence?* The Darwin College Lectures. Cambridge, England: Cambridge University Press.

De Renzi, E., Faglioni, P., Savoiardo, M., & Vignolo, L. A. (1966). The influence of aphasia and of the hemispheric side of the cerebral lesion on abstract thinking. *Cortex, 2,* 399–420.

De Renzi, E., Faglioni, P., Scotti, G., & Spinnler, H. (1972). Impairment of color sorting behaviours after hemispheric damage: An experimental study with the Holmgren skein test. *Cortex, 8,* 147–163.

De Renzi, E., Spinnler, H., Scotti, G., & Faglioni, P. (1972). Impairment in associating color to form, concomitant with aphasia. *Brain, 95,* 293–304.

Druks, J., & Shallice, T. (2000). Selective preservation of naming from description and the "restricted preverbal message." *Brain and Language, 72,* 100–128.

Enns, J. T., & Di Lollo, V. (2000). What's new in visual masking? *Trends in Cognitive Sciences, 4,* 345–352.

Erdfelder, E., & Bredenkamp, J. (1998). Recognition of script-typical versus script-atypical information: Effects of cognitive elaboration. *Memory & Cognition, 26,* 922–938.

Fodor, J. (1975). *The language of thought.* Cambridge, MA: Harvard University Press.

Franks, J. J., & Bransford, J. D. (1971). Abstraction of visual patterns. *Journal of Experimental Psychology, 90,* 65–74.

Gainotti, G., Carlomagno, S., Craca, A., & Silveri, M. C. (1986). Disorders of classificatory activity in aphasia. *Brain and Language, 28,* 181–195.

Gauthier, I., James, T. W., Curby, K. M., & Tarr, M. J. (2003). The influence of conceptual knowledge on visual discrimination. *Cognitive Neuropsychology, 20,* 507–523.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition, 23,* 183–209.

Gibson, J. J., & Gibson, E. J. (1955). Perceptual learning: Differentiation or enrichment. *Psychological Review, 62,* 32–41.

Gleitman, L., & Papafragou, A. (2005). Language and thought. In K. Holyoak & B. Morrison (Eds.), *Cambridge handbook of thinking and reasoning* (pp. 633–661). Cambridge, England: Cambridge University Press.

Goldstein, K. (1948). *Language and language disturbances.* New York: Grune & Stratton.

Goldstein, K. (1971). The problems of the meaning of words based upon observation of aphasic patients. In A. Gurwitsch, E. M. Goldstein Haudek, & W. E. Haudek (Eds.), *Selected papers: Phainomenologica* (Vol. 43, pp. 344–359). The Hague, the Netherlands: Nijhoff. (Original work published 1936)

Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition: As soon as you know it is there, you know what it is. *Psychological Science, 16,* 152–160.

Gronlund, S. D., & Ratcliff, R. (1989). Time course of item and associative information: Implications for global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 846–858.

Hampton, J. A. (2006). Concepts as prototypes. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 46, pp. 79–113). London: Elsevier.

Hanawalt, N. G., & Demarest, I. H. (1939). The effect of verbal suggestion in the recall period upon the reproduction of visually perceived forms. *Journal of Experimental Psychology, 25,* 159–174.

Harnad, S. (2005). Cognition is categorization. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization* (pp. 20–42). Amsterdam: Elsevier.

He, Z. J., & Nakayama, K. (1992, September 17). Surfaces versus features in visual search. *Nature, 359,* 231–233.

Hirshman, E., Trembath, D., & Mulligan, N. (1994). Theoretical implications of the mnemonic benefits of perceptual interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 608–620.

Hjelmquist, E. K. E. (1989). Concept-formation in non-verbal categorization tasks in brain-damaged patients with and without aphasia. *Scandinavian Journal of Psychology, 30,* 243–254.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review, 98,* 352–376.

Kail, R., & Nippold, M. A. (1984). Unconstrained retrieval from semantic memory. *Child Development, 55,* 944–951.

Kapur, S., Craik, F. I. M., Tulving, E., Wilson, A. A., Houle, S., & Brown, G. M. (1994). Neuroanatomical correlates of encoding in episodic memory: Levels of processing effect. *Proceedings of the National Academy of Sciences, USA, 91,* 2008–2011.

Katz, P. A. (1963). Effects of labels on children's perception and discrimination: Learning. *Journal of Experimental Psychology, 66,* 423–428.

Kelter, S., Cohen, R., Engel, D., List, G., & Strohner, H. (1976). Aphasic disorders in matching tasks involving conceptual analysis and covert naming. *Cortex, 12,* 383–394.

Kelter, S., Cohen, R., Engel, D., List, G., & Strohner, H. (1977). Conceptual structure of aphasic and schizophrenic patients in a nonverbal sorting task. *Journal of Psycholinguistic Research, 6,* 279–303.

Koemeda-Lutz, M., Cohen, R., & Meier, E. (1987). Organization of and access to semantic memory in aphasia. *Brain and Language, 30,* 321–337.

Koutstaal, W., Reddy, C., Jackson, E. M., Prince, S., Cendan, D. L., & Schacter, D. L. (2003). False recognition of abstract versus common objects in older and younger adults: Testing the semantic categorization account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 499–510.

Koutstaal, W., & Schacter, D. L. (1997). Gist-based false recognition of pictures in older and younger adults. *Journal of Memory and Language, 37,* 555–583.

Kuhl, P. (1994). Learning and representation in speech and language. *Current Opinions in Neurobiology, 4,* 812–822.

Landau, B., & Shipley, E. (2001). Labelling patterns and object naming. *Developmental Science, 4,* 109–118.

Li, P., & Gleitman, L. (2002). Turning the tables: Language and spatial reasoning. *Cognition, 83,* 265–294.

Lockhart, R. S., Craik, F. I. M., & Jacoby, L. L. (1976). Depth of processing, recognition and recall: Some aspects of a general memory system. In J. Brown (Ed.), *Recall and recognition* (pp. 75–102). London: Wiley.

Loewenstein, J., & Gentner, D. (1998). Relational language facilitates analogy in children. *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society, 20,* 615–620.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science, 18,* 1077–1083.

Markman, E. M., & Hutchinson, J. E. (1984). Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive Psychology, 16,* 1–27.

Marks, W. (1991). Effects of encoding the perceptual features of pictures on memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 566–577.

McClelland, J. L. (1995). Constructive memory and memory distortions: A parallel-distributed-processing approach. In D. L. Schacter (Ed.), *Memory distortion* (pp. 69–90). Cambridge, MA: Harvard University Press.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105,* 724–760.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88,* 375–407.

Melcher, J. M., & Schooler, J. W. (1996). The misremembrance of wines past: Verbal and perceptual expertise differentially mediate verbal overshadowing of taste memory. *Journal of Memory and Language, 35,* 231–245.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16,* 519–533.

Musen, G. (1991). Effects of verbal labeling and exposure duration on implicit memory for visual patterns. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 954–962.

Pederson, E., Danziger, E., Wilkins, D., Levinson, S., Kita, S., & Senft, G. (1998). Semantic typology and spatial conceptualization. *Language, 74,* 557–589.

Pinker, S. (1994). *The language instinct: How the mind creates language.* New York: William Morrow.

Postman, L. (1955). Association theory and perceptual learning. *Psychological Review, 62,* 438–446.

Prentice, W. C. H. (1954). Visual recognition of verbally labeled figures. *American Journal of Psychology, 67,* 315–320.

Roberson, D. (2006). How language helps category acquisition. *The 28th Annual Conference of the Cognitive Science Society* (p. 2660). Mahwah, NJ: Erlbaum.

Roberson, D., Davidoff, J., & Braisby, N. (1999). Similarity and categorisation: Neuropsychological evidence for a dissociation in explicit categorisation tasks. *Cognition, 71,* 1–42.

Robinson, J. S. (1955). The effect of learning verbal labels for stimuli on their later discrimination. *Journal of Experimental Psychology, 49,* 112–114.

Roediger, H. L. I., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 803–814.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 111–144). New York: Academic Press.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Erlbaum.

Schooler, J. W. (2002). Verbalization produces a transfer inappropriate processing shift. *Applied Cognitive Psychology, 16,* 989–997.

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

Shiffrin, R. M., & Steyvers, M. (1997). Model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4,* 145–166.

Sloutsky, V. M., & Fisher, A. V. (2004a). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General, 133,* 166–188.

Sloutsky, V. M., & Fisher, A. V. (2004b). When development and learning decrease memory: Evidence against category-based induction in children. *Psychological Science, 15,* 553–558.

Sloutsky, V. M., & Lo, Y. F. (1999). How much does a shared name make things similar? Part 1. Linguistic labels and the development of similarity judgment. *Developmental Psychology, 35,* 1478–1492.

Sloutsky, V. M., Lo, Y. F., & Fisher, A. V. (2001). How much does a shared name make things similar? Linguistic labels, similarity, and the development of inductive inference. *Child Development, 72,* 1695–1709.

Smith, L. B., Jones, S. S., & Landau, B. (1996). Naming in young children: A dumb attentional mechanism? *Cognition, 60,* 143–171.

Spencer, T. J., & Shuntich, R. (1970). Evidence for an interruption theory of backward masking. *Journal of Experimental Psychology, 85,* 198–203.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs, 74,* 1–29.

Suzuki, S., & Cavanagh, P. (1995). Facial organization blocks access to low-level features: An object inferiority effect. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 901–913.

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

Waxman, S. R., & Hall, D. G. (1993). The development of a linkage between count nouns and object categories: Evidence from 15-month-old to 21-month-old infants. *Child Development, 64,* 1224–1241.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology, 29,* 257–302.

Wickham, L. H. V., & Swift, H. (2006). Articulatory suppression attenuates the verbal overshadowing effect: A role for verbal encoding in face identification. *Applied Cognitive Psychology, 20,* 157–169.

Woodward, A. E., Bjork, R. A., & Jongewar, R. H. (1973). Recall and recognition as a function of primary rehearsal. *Journal of Verbal Learning and Verbal Behavior, 12,* 608–617.

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition, 85,* 223–250.

Received May 27, 2007
Revision received November 7, 2007
Accepted November 7, 2007 ■