Cognitive, Affective, & Behavioral Neuroscience
2008, 8 (1), 3-16
doi: 10.3758/CABN.8.1.3

The case for implicit category learning

Edward E. Smith
Columbia University, New York, New York

This article evaluates the evidence regarding the claim that people can learn a novel category implicitly—that is, by an implicit memory system that is qualitatively different from an explicit system. The evidence that is considered is based on the prototype extraction task, in which participants are first exposed to a set of category exemplars under incidental learning instructions and are then required to categorize novel test items. Knowlton and Squire (1993) first reported that memory-impaired patients performed normally on the prototype extraction task while being impaired on a comparable recognition task. Several studies have replicated these results, but other articles have criticized the evidence for implicit category learning on both methodological and theoretical grounds. In this article, we consider five of these criticisms—for example, that the normal performance of the patients is due to intact working memory mechanisms (see, e.g., Palmeri & Flannery, 1999) or to the lesser cognitive demands of prototype extraction rather than recognition (e.g., Nosofsky & Zaki, 1998). For each of the five criticisms, we offer counterevidence that supports implicit category learning.

Psychologists and neuroscientists began to recognize the existence of an implicit memory system during the 1960s and early 1970s when they discovered that a medial temporal lobe (MTL) amnesic, H.M., could learn a new perceptual motor skill as readily as neurologically intact participants. Soon after, researchers discovered that MTL patients showed intact perceptual priming, which indicated that implicit memory is not confined to motor skills (see, e.g., Schacter, 1987, and Squire, 2004, for reviews of these early developments). These findings led to a characterization of implicit memory as a system that can use past experience without intention or awareness and that does not depend on MTL.

In 1993, Knowlton and Squire published the first article claiming to demonstrate that memory-impaired patients had an intact ability to learn a novel category. This is known as *implicit category learning*. These findings were important because they demonstrated a kind of implicit learning that extracts common features and supports generalizations to novel items, which seemed to go beyond perceptual priming or other known forms of implicit learning. However, in recent years, the evidence for implicit category learning has come under sharp criticism, primarily from researchers who have focused on behavioral and mathematical analyses of category learning (e.g., Nosofsky & Zaki, 1998; Zaki & Nosofsky, 2001). The purpose of this article is to review and evaluate the evidence for and against postulating an implicit category-learning system, particularly in light of recent findings on this topic (e.g., Bozoki, Grossman, & Smith, 2006; Reber, Gitelman, Parrish, & Mesulam, 2003). The argument will be made that these recent articles provide new evidence that (1) people can indeed learn a category implicitly, (2) patients with compromised MTL function learn roughly as well as normals, and (3) the learning involved may be a kind of perceptual fluency.

Initial Evidence for Implicit Category Learning

In their seminal study, Knowlton and Squire (1993) used a *prototype extraction* task. Both memory-impaired patients and normal controls were presented a series of 40 dot patterns (see Figure 1A for examples). All of the patterns had been created by starting with a prototype pattern and then transforming it to some degree by moving some proportion of the dots (after Posner & Keele, 1968). During training, nothing was mentioned about a category, since both patients and controls were instructed simply to point to the center dot in each pattern. After training, all participants were informed that the patterns they had just seen "all belonged to a single category" and that they now had to determine which of a sequence of test patterns also belonged to that category. This procedure is standard implicit memory methodology: The purpose of the training information was disguised so that participants did not intentionally try to remember it; hence, they presumably had to rely on an implicit system during test. During the test phase, a total of 80-plus novel items were presented, which varied in how similar they were to the prototype that had spawned the training items. Both the patients and controls performed the unexpected categorization task with above-chance accuracy, and the patients performed as accurately as the controls. These results were in sharp contrast to findings obtained in a test of recognition memory for the same kind of dot patterns. In the latter task, the patients performed significantly worse than did the controls, although the test was not particularly demanding (the study items consisted of five different dot patterns presented eight times each).[1]

E. E. Smith, eesmith@psych.columbia.edu

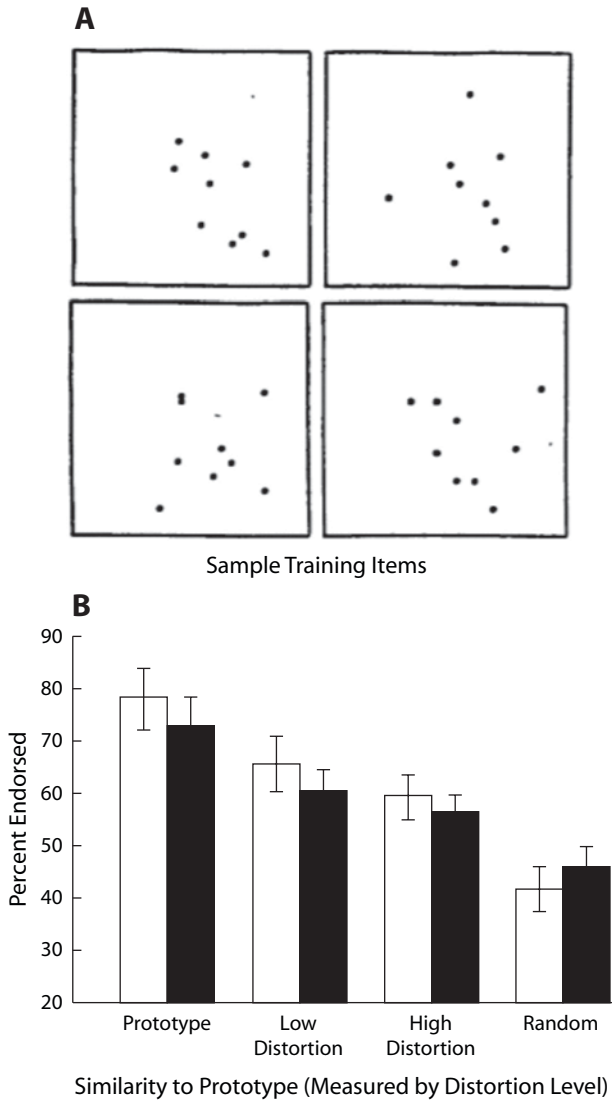

Figure 1. Knowlton and Squire (1993) experiment. Panel A contains examples of the items used during training. Panel B presents, in bar-graph form, prototypicality gradients obtained from test trials, for both memory-impaired patients and matched controls. A prototype test item is, of course, maximally similar to the prototype, a low-distortion test item somewhat less similar, a high-distortion item even less similar, and a random test item is the least similar to the prototype. The black bars depict the data for the patients, the white bars the data for controls. From "The Learning of Categories: Parallel Brain Systems for Item Memory and Category Knowledge," by B. J. Knowlton and L. R. Squire, *Science*, *262*, p. 1748. Copyright 1993 by the American Association for the Advancement of Science. Adapted with permission.

Further evidence for implicit category learning was provided by examining the extent to which a participant endorsed a test item as a category member as a function of its similarity to the underlying prototype of the category (which was not presented during training). This *prototypicality gradient* is presented in Figure 1B, in which more extreme values on the *x*-axis represent less similarity to the prototype. There appears to be little difference between the memory-impaired patients and controls in their gradients.

The conclusion that was drawn by Knowlton and Squire (1993) was that both groups of participants had learned the category implicitly rather than explicitly. Several subsequent experiments have used a variant of this prototype extraction task and have replicated the findings above for patients—intact categorization, but impaired recognition—using as patients either MTL amnesics (Kolodny, 1994; Reed, Squire, Patalano, Smith, & Jonides, 1999; Squire & Knowlton, 1995) or people with mild Alzheimer's disease (Bozoki et al., 2006; for partial replications, see also Keri, Kalman, Kelemen, Benedek, & Janka, 2001; Keri et al., 1999).

The Reed et al. (1999) study made an additional important point. In this experiment, the items that were used were not abstract dot patterns, but a set of artificial animals (Figure 2A illustrates some of the animals). Whereas abstract dot patterns are difficult to describe, Reed et al.'s artificial animals are not, since they are composed of salient and separable dimensions—for example, a creature with a round, striped body and a long neck. The fact that Reed et al. obtained the usual results for memory-impaired patients with these items—intact categorization with impaired recognition, and a roughly normal prototypicality gradient for categorization—suggests that the implicit categorization system of interest operates on both easy- and difficult-to-describe materials.

The prototype extraction task shares certain features with other paradigms that are used to study implicit memory. The training phase contains no feedback and is disguised as something other than a learning task; the purpose of the disguise is to block intentional learning and the subsequent use of explicit memory. This aspect distinguishes prototype extraction from two other tasks that have been used to study implicit category learning. One is *probabilistic categorization*, which has been used extensively in neuropsychological studies (e.g., Knowlton, Mangels, & Squire, 1996; Shohamy et al., 2004). In this paradigm, a set of one to three different visual patterns is presented on each trial, and the participant has to decide in which of two categories each set belongs. Unlike the prototype extraction task, in probabilistic categorization, feedback is given during the training phase, and the identical stimulus can be assigned to different categories on different trials. In view of these differences, it is not surprising that performance on the probabilistic-categorization and prototype extraction tasks can be dissociated: Patients with Parkinson's disease are impaired on probabilistic categorization, but not on prototype extraction (Ashby & Ell, 2001; Keri, 2003).

There are also neuroimaging results that imply that probabilistic categorization is based on a system different than that mediating prototype extraction. When young normals are imaged while performing the probabilistic-categorization task, the striatum is consistently activated (see, e.g., Poldrack et al., 2001), and differences in striatal activation are correlated with differences in behavioral performance (Foerde, Knowlton, & Poldrack, 2006). In contrast, the striatum seems to play little role when young normals are imaged while performing prototype extraction (e.g., Reber et al., 2003).

The other implicit category learning task of note is the *information integration* task that has been used extensively by Ashby and Maddox and their colleagues (see, e.g., Ashby & Maddox, 2005, for a recent review). In a typical version, a two-dimensional stimulus is presented, and participants must decide to which of two categories the stimulus belongs. Feedback is given during the training phase, and the stimuli are constructed so as to be difficult to describe. These characteristics are unlike those of prototype extraction; as already noted, Reed et al. (1999) used materials that were relatively easy to articulate in prototype extraction and obtained the standard results (see also Bozoki et al., 2006). Again, the two tasks dissociate: Parkinson's patients are impaired on information integration, but not on prototype extraction (Ashby & Ell, 2001). Furthermore, in one of the only neuroimaging studies to use an information integration task, activation of the striatum was again a major finding (Seger & Cincotta, 2002). Thus, the two alternative category-learning paradigms seem to be tapping a system different from that recruited by prototype extraction, and we will not dwell on them in what follows.

Criticisms of Prototype Extraction Research

Several criticisms of the prototype extraction research that supports implicit category learning have been made, and one way to appreciate them is through the lens of memory systems. The research at issue assumes that results in the prototype extraction task reflect only implicit memory, with no contribution of explicit memory or working memory, and this is presumably the case for both memory-impaired patients and controls. All of the criticisms essentially challenge this assumption and argue, for example, that performance on prototype extraction by memory-impaired patients is entirely due to either explicit memory (see, e.g., Nosofsky & Zaki, 1998), or working memory (e.g., Palmeri & Flanery, 1999). More specifically, there are five criticisms of concern, with some being made on the basis of empirical findings and others on the basis of the results of computational modeling.

1. The seemingly intact performance of memory-impaired patients on prototype extraction can be attributed entirely to the patients' relying on working memory mechanisms during the test trials (see, e.g., Palmeri & Flanery, 1999).

2. The seemingly intact performance of memory-impaired patients on prototype extraction in at least one particular study—Reed et al. (1999)—can be explained (modeled) by assuming that participants need explicitly remember information about only one or two features, acquired during either the training or the test phase (Zaki & Nosofsky, 2001).

3. The performance of memory-impaired patients on prototype extraction is not truly intact, as revealed by meta-analysis techniques. This means that the performance of memory-impaired patients on categorization and their performance on recognition are only quantitatively dissimilar, which suggests that performance on both tasks may be based on the same system (Zaki, 2004).

4. The findings that memory-impaired patients are intact on prototype extraction but impaired on a recognition task is due to the categorization tasks being less difficult than the recognition task (Zaki, Nosofsky, Jessup, & Unverzagt, 2003).

5. Intact performance of memory-impaired patients on prototype extraction, along with their impaired performance on a standard recognition memory task, can be explained (modeled) entirely in terms of retrieval of exemplars from explicit memory (see, e.g., Nosofsky & Zaki, 1998).[2]

The gist of the criticisms is this. Most researchers would readily acknowledge that there are two different category-learning systems: an exemplar retrieval system based on explicit memory that determines the similarity of a test item to stored exemplars of a category and uses this as a guide to categorization, and a system based on working memory that tests hypotheses about what features are relevant to category membership and uses these features to guide categorization (Ashby & O'Brien, 2005; Smith, Patalano, & Jonides, 1998). But there is no convincing evidence for the existence of a third category-learning system, one based on implicit memory. In the next section, we will consider in detail each of these criticisms, along with rebuttals to them.

EVALUATION OF THE EVIDENCE

Are Results in the Prototype Extraction Task Due to Working Memory?

Using the standard prototype extraction task, Palmeri and Flanery (1999) demonstrated that one can learn a category solely from the test trials. In an ingenious experiment, the participants (young normals) were misleadingly told that a series of dot patterns had been presented to them subliminally (while they were doing another task) and that these subliminal patterns were all instances from the same category. The participants were further informed that they now had to determine which of a sequence of test patterns also belonged to the category. Essentially, the paradigm was the same as that of Knowlton and Squire (1993), except that no training patterns were ever presented. Nevertheless, the participants in Palmeri and Flanery performed the categorization task with above-chance accuracy and showed a prototypicality gradient like that in Figure 1B (i.e., the probability of endorsing a test item as a category member decreased as the similarity of the test item to the category prototype decreased). The undeniable conclusion from these results was that category learning had occurred during the test trials (there were 80 such trials).[3]

More to the point, the evidence that previous investigators had used to support implicit category learning during training now could be attributed to learning during the test trials, and the latter could have involved working memory, which is known to be relatively intact in some memory-impaired patients. More specifically, knowing that a category is present, and confronted with the first few test trials, participants may have tested hypotheses about which features were diagnostic of the category (e.g.,

with dot patterns, participants might have noted that most of the first few test patterns contained a cluster of dots in their upper left-hand quadrants; with artificial animals, participants might have noticed that most of the first few test items contained striped creatures with long necks). On each successive test trial, participants might have checked for these critical features or, if too many patterns had them, sought another feature set that would have allowed them to sometimes declare an item a nonmember. In short, participants could have engaged in hypothesis testing, a means of category learning that is heavily dependent on working memory, not explicit memory, and that is primarily mediated by prefrontal structures, not MTL ones (see, e.g., Ashby & O'Brien, 2005). In addition to its plausibility, this working memory proposal has some empirical support. When schizophrenic patients were tested in the no-training variant of the prototype extraction task, those patients who were known in advance to have intact working memory performed normally on the task, whereas those patients with a known deficiency in working memory were impaired on the task (see Keri, 2003).[4]

The fact that participants in standard prototype extraction tasks (i.e., those that include training trials) *could* have learned the category solely on test trials does not mean that they *did* learn it in this fashion. To determine how participants actually do learn in a standard paradigm, we need a direct comparison of performance on prototype extraction with and without training trials: Performance without training trials may be used as a rough estimate of what can be learned during test via working memory alone, and this estimate can be subtracted from performance with training trials to determine whether there is residual learning that can be attributed to an implicit system.

Bozoki, Grossman, and Smith (2006) recently reported an experiment using this kind of subtraction logic. The participants were patients with mild Alzheimer's disease (who had extensive damage in MTL and, thus, compromised explicit memory) and age-matched normal controls. The task was prototype extraction, with the items being the artificial animals used by Reed et al. (1999); see Figure 2A. On the measure of overall categorization accuracy, the Alzheimer's participants showed significant residual learning after subtracting their working memory component, and their implicit category learning was indistinguishable from that of normal controls. The prototypicality gradients for patients with and without training items are presented in Figure 2B. (Again, more extreme values on the *x*-axis represent less similarity to the prototype.) Clearly, patients with impaired memory learned something about the category during the training trials, and what they learned did not differ significantly from what controls learned; that is, there was no difference between the prototypicality gradients for the AD patients and controls. In sum, these results offset the working memory criticism and provide some evidence for implicit category learning.[5]

Will these results for patients generalize to other kinds of materials, particularly the dot patterns that have figured so prominently in this research? Earlier, we noted that the basic phenomena obtained in studies with dot patterns—intact categorization with impaired recognition and a roughly normal prototypicality gradient—are obtained with artificial animals as well (see the introduction). But these past correspondences do not guarantee that the results in Figure 2B will also generalize across materials. And there is reason to be cautious. The Palmeri and Flanery (1999) study, which used dot patterns and introduced the no-training condition, also included a standard condition in which normal participants were exposed to training items before being tested. The authors found no beneficial effect of training. This null result contrasts with the positive one of Bozoki et al. (2006), and the discrepancy may well reflect the difference in materials, dot patterns being substantially less analyzable and verbalizable than artificial animals. The discrepancy in results may also partly reflect the difference in participants—older, memory-impaired patients in Bozoki et al. (2006), as opposed to young normals in Palmeri and Flanery (1999). Normals are more likely than patients to be strategic about their use of working memory during test, deploying this system more when they have had no training than when they have had training and learned something from it (see Bozoki et al., 2006). This kind of strategic control in normals would compromise the subtraction logic of the experiments of interest.

What is needed are experiments with memory-impaired patients that contrast training and no-training conditions with both artificial animals and dot patterns. As things stand now, the Bozoki et al. (2006) results provide evidence that memory-impaired patients are capable of implicit category learning, at least with analyzable materials.

Do Participants Explicitly Remember Only a Few Features?

A number of researchers have proposed that in some of the studies supporting implicit category learning, participants may have attended to only one or two relatively salient features of the patterns rather than formed a representation of the entire pattern. This criticism has been most developed in Zaki and Nosofsky's (2001) critique of the Reed et al. (1999) study that used artificial animals as items.

Zaki and Nosofsky (2001) consider two different versions of this criticism. (1) Using explicit memory, participants need remember only one or two salient features of the training items (which presumably is within the capacity of memory-impaired patients) and then employ these features as a basis for categorization during test, and (2) using working memory, during the test phase, participants discover one or two critical features and employ them as a basis for categorization decisions. The second version is easier to deal with. It is just a special case of the previous criticism, and again, the Bozoki et al. (2006) findings—evidence for category learning after test phase learning has been subtracted—provide a counterargument to the criticism.

The first version of the criticism requires more extended discussion. Reed et al. (1999) did, in fact, consider the hypothesis that their patients made their categorizations on the basis of just 1 or 2 of the 10 relevant features and tried

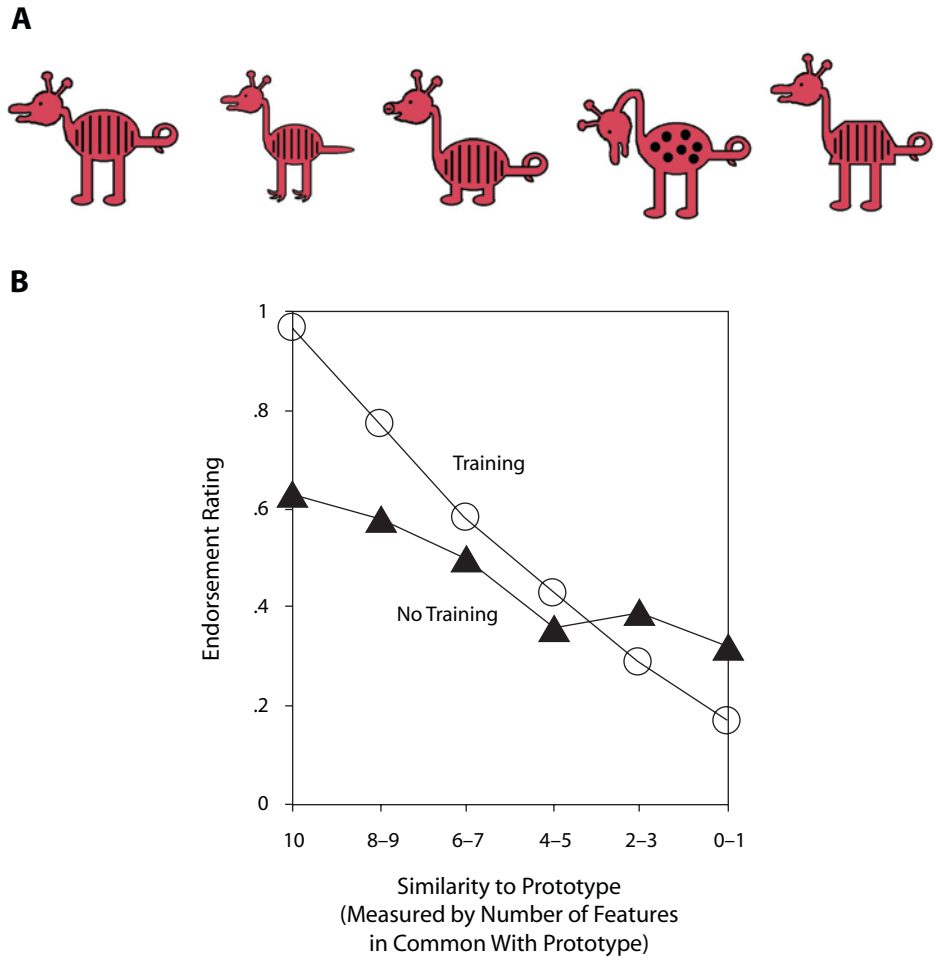

Figure 2. Bozoki, Grossman, and Smith (2006) experiment. Panel A contains examples of the items used. Panel B presents prototypicality gradients for Alzheimer's patients who did or did not receive training items (white circles and black triangles, respectively). Importantly, the gradient for patients who received training items is steeper than that for patients who had no training items, and steepness is a measure of learning. Not endorsing a test item that shares fewer than four features with the prototype is considered a correct response. From "Can Patients With Alzheimer's Disease Learn a Category Implicitly?" by A. Bozoki, M. Grossman, and E. E. Smith, *Neuropsychologia*, *44*, p. 822. Copyright 2006 by Elsevier B.V. Adapted with permission.

to evaluate this hypothesis in a couple of ways. As one example, for each participant, they determined the feature, or attribute value, that was the most frequently endorsed in order to see whether that endorsement rate was close to 100% (as it might be if participants made their categorizations solely on the basis of 1 feature). Reed et al. found that on average, the most frequently endorsed feature was endorsed only 77% of the time, and they offered this result as evidence against the hypothesis that patients used only a single feature during categorization. In a more recent study (Koenig et al., 2007) that used prototype extraction with artificial animals that varied on six attributes, we performed similar tests of the hypothesis of interest. For each of 23 mild AD patients and a like number of age-matched controls, we asked whether there was any attribute that had one of its values endorsed 80% of the time and the other value endorsed less than 20% of the time. Only 2 of the 23 patients showed this pattern of single-feature use (none of the controls did), and when we removed these 2 patients from our data analysis, there was still evidence for implicit category learning in the patient group.

Zaki and Nosofsky (2001) raised a problem with such straightforward behavioral measures. Even when participants are using only one feature, they may use a probabilistic response rule (e.g., "If most of the training items were red, then endorse a red test item 75% of the time") rather than a deterministic rule (e.g., "Endorse red items 100%"). One way to deal with this potential problem is to use a mathematical model of category learning that includes a parameter for the response rule, as well as parameters for other relevant processes. In doing so, one can see how well the model fits data for memory-impaired patients under the assumption that only one or two features are attended and the response rule is probabilistic. Zaki and Nosofsky approximated this strategy in an experiment that was patterned on the Reed et al. (1999) study, and

they found that an attend-to-one-or-two-features version of the model provided a good fit to their data.

However, their research is problematic in two respects. First, rather than testing patients and controls, Zaki and Nosofsky (2001) tested only young normals, with one group of participants receiving the test immediately after training (the immediate group), and the other group receiving the test after a 1-week delay (the delay group). The delay group is a "stand-in" for memory-impaired patients. This is a rather pale approximation to memory impairment caused by brain damage (as the authors themselves note). When Alzheimer's patients are presented a list of words to remember and are tested on them 10 min later (a standard procedure in some Alzheimer's clinics), not only do the patients fail to recall any of the words, often they cannot remember ever being tested (Grossman, 2007, personal communication). It seems unlikely that the young normals in Zaki and Nosofsky's delay condition would forget that they were given a memory test. Thus, although Zaki and Nosofsky's results show that young normals may perform the test phase of prototype extraction by recalling one or two features from a training session a week ago, it is by no means clear that memory-impaired patients could recall this much information from material presented a few minutes ago.

Second, Zaki and Nosofsky's (2001) means of determining the amount of attention paid to a particular attribute was quite removed from the actual data. In the first step of their analysis, the attentional value, or "weight," of an attribute was determined by a model that included 11 free parameters for every participant, with a separate parameter for each relevant attribute. The authors found that attention weights for only one to two of these attributes were needed to account for the participants' categorizations. Because there are so many free, unconstrained parameters, the results do not seem that compelling. In the second step of their analysis, Zaki and Nosofsky fit more restricted versions of their model that contained between one and three attention-weight parameters, and they showed that these restricted models provided fits to the categorization data that were as good as those of the full model (the 11-parameter version). This seems a more impressive result. Note, however, that the restricted models were based on the full, 11-parameter version; for example, the single attribute used in a one-parameter restricted model was the attribute with the greatest attention weight in the full model.

We are thus left with a conclusion—categorization is based on one or two explicitly remembered features—that is based on an analytic framework that has a large number of unconstrained parameters and that has never been applied to memory-impaired patients. Further applications of the model may lead to a more convincing argument, particularly if it is applied to patients. But the results that were just discussed do not seem strong enough to overthrow a conclusion made on the basis of direct behavioral measures of patients' performances on prototype extraction. We have not reached a resolution of this issue.

Is the Performance of Memory-Impaired Patients in Prototype Extraction Truly Intact?

Every prototype extraction study that has compared normal controls and memory-impaired patients has shown that controls do slightly better, even though this difference has failed to reach statistical significance. Although the differences may have been small in each study, their consistency across studies led Zaki (2004) to perform some meta-analyses of category-learning tasks that involved memory-impaired patients. She analyzed over 14 different experiments, 5 of which used the prototype extraction task. These analyses showed that controls scored significantly higher on implicit category learning than did memory-impaired patients. This finding indicates that the performance of memory-impaired patients on categorization is only quantitatively dissimilar from their performance on recognition (they are impaired on both), which suggests that both tasks may recruit the same system. Zaki used this finding in conjunction with a model-based analysis to argue that the only system involved in prototype extraction is explicit memory. In essence, although memory-impaired patients have reduced memory sensitivity (in a sense that will soon be described in detail), it is sufficient to score close to normal on the categorization task, but not on a recognition task.

From the present perspective, a problem with Zaki's (2004) meta-analysis is that it mixes together studies using different category-learning paradigms—including prototype extraction, probabilistic-categorization, and information integration tasks—and we have already noted that the latter two paradigms likely reflect different mechanisms than does prototype extraction (see the introduction). However, even if we assume that Zaki's finding of a small advantage for controls applies to prototype extraction and further grant the claim that this advantage is due to controls, having better explicit memory, this does not necessarily imply that the patients in these studies also used explicit memory. Indeed, this conclusion seems particularly implausible for the results obtained by Squire and Knowlton (1995) and Reed et al. (1999). In both these studies, one of the patient participants, E.P., performed at chance on all measures of explicit memory, but normally on implicit category learning, which undermines Zaki's conclusion to some degree. If category learning is due only to explicit memory and E.P. has no discernible explicit memory, how can he perform normally on category learning? (This same point has been discussed by Knowlton, 1999, and Nosofsky & Zaki, 1999.) In short, with respect to explaining the results of neuropsychological experiments, Zaki's argument has parsimony on its side, but it lacks plausibility.

An alternative proposal to Zaki (2004) is that during the training period, both patients and controls acquired information only implicitly, but during the test period, the controls, but not the patients, also tried to explicitly recall training patterns to help guide their categorization decisions. This alternative acknowledges that prototype extraction can involve explicit memory but restricts this

involvement to people who have neurologically intact memory systems.

Interestingly, the point of agreement between the two proposals—that an implicit task recruits an explicit system in normal participants—is rarely considered in the relevant experimental reports. Typically, researchers do not ask participants whether they tried to remember training instances during the test phase. Such postexperimental interviews also appear to be rare in other studies of implicit learning, such as repetition priming. In a recent study of repetition priming in younger and older normal participants that did ask such questions, fully one third of all participants volunteered that they had indeed tried to recall the training items during the test phase (May, Hasher, & Foong, 2005).[6]

Is the Difference in Performance of MTL Patients on Categorization and Recognition Tasks Due to Task Difficulty?

We have been focusing on whether implicit categorization is really intact, but it is important to keep in mind that the phenomena of interest include impaired recognition along with intact categorization. In a recent attempt to account for this dissociation, Zaki, Nosofsky, Jessup, and Unverzagt (2003) suggested that the categorization tasks that have been used with memory-impaired patients are simply less demanding than the recognition tasks used. Taken at face value, this claim is at odds with certain basic findings. The standard way of determining the relative difficulty of two tasks is to see which task leads to poorer performance in normal controls. By this standard, Knowlton and Squire's (1993) categorization task was more difficult than their recognition task, since the percent correct on the two tasks was roughly 60% and 80%, respectively. (This difference likely reflected the fact that Knowlton and Squire used 40 different training items in categorization, but only 5 different study items in recognition.) Indeed, even if we consider only the performance of Knowlton and Squire's patients, their performance on recognition is no worse than their performance on categorization. There is simply no evidence in this study (or Squire & Knowlton's, 1995, follow-up experiments) to support the claim that the dissociation is due to differential difficulty. Squire and Knowlton (2000) have argued roughly the same point.

The only way to salvage the differential-difficulty hypothesis is by focusing on a component process of the task, not the whole task, and to argue that this process is more taxed in the recognition than in the categorization conditions in studies that have shown the dissociation. Zaki et al. (2003) take this tack as well, and they argue that the critical process involves discriminating between memory representations. They further propose that such discrimination is impaired in memory-impaired patients (essentially, any comparison between a test item and a memory representation is degraded by noise), and that prototype extraction requires less discrimination sensitivity than do the recognition tasks used. Note that this kind of differential difficulty argument rests on a computational analysis of the tasks, not on empirical data about the overall difficulty of the task. And note further that even if recognition taxes a memory discrimination process more than categorization does, it does not follow that recognition should be the more difficult task. Among other considerations, there are processes in addition to memory discrimination that contribute to performance in the two tasks (like the response rule that we considered earlier), and some of these additional processes may favor recognition over categorization.

To provide data on the point, Zaki et al. (2003) compared memory-impaired patients and controls on two categorization tasks that presumably differed in the demand they made on memory discrimination. One task was standard prototype extraction—it was exactly the Knowlton and Squire (1993) dot-pattern paradigm. The other task required participants to learn two categories of dot patterns concurrently, the two categories being generated from different prototypes. On each training trial, a pattern would appear; the participant would indicate whether it belonged to Category A or B and then would receive feedback. During the test trials, participants had to categorize novel items (as well as the old training items). As was expected, the patients performed less well than the controls, at least in the final blocks of trials (55% vs. 63%).

This dual-category task has little connection to the kind of implicit category learning that is at issue in this article. For one thing, participants were instructed during training that categories were present; for another, they received feedback. For both these reasons, participants presumably had an intention to learn, which should have led them (at least the control participants) to apply explicit memory processes. All this study shows is that in a task that recruits and requires explicit memory, patients with memory impairment perform worse than controls. We already knew this from studies of recognition memory.

In sum, there is no merit to the global claim that the prototype extraction task is easier than the recognition tasks with which it has been compared. However, there may be merit to the idea that a common component of the categorization and recognition tasks is more taxed in the recognition task (see the next subsection), but the experiment provided as support for this idea by Zaki et al. (2003) does not seem to deal with implicit category learning.

Can the Performance of Memory-Impaired Patients on Prototype Extraction and Recognition Be Explained (Modeled) Entirely in Terms of Explicit Memory?

Basic argument. In their initial critique of the research on implicit category learning, Nosofsky and Zaki (1998) argued that the Knowlton and Squire (1993) findings on categorization and recognition could be explained in terms of a single-system, explicit-memory model. The findings at stake are the dissociation between recognition and categorization, and the prototypicality gradient for categorization. The model at issue, the generalized context model

(or GCM), is a mathematical model of exemplar-based categorization that was developed by Nosofsky (see, e.g., 1988, 1991). According to the GCM, during the training phase in prototype extraction experiments, participants store a representation of each exemplar presented to them; during the test phase, they determine whether a test item belongs to the category by determining its similarity to each exemplar, summing these similarities, and comparing the summed similarity with a categorization criterion. Exactly the same processes go on in a recognition task in determining whether a test item is old or new, except that the summed similarities are now compared with a recognition criterion that need not be the same as the categorization criterion.

Nosofsky and Zaki (1998) obtained empirical estimates of the exemplar similarities involved and used them to fit a multiparameter version of the GCM to the Knowlton and Squire (1993) data. To account for differences between the memory-impaired patients and controls, Nosofsky and Zaki (1998) altered one parameter of their model—that which reflects memory discrimination, in the sense that less discrimination means that the outcome of a similarity comparison is degraded. The memory discrimination parameter was constrained to be lower for memory-impaired patients than for controls. This model was shown to provide an excellent fit to both sets of critical findings: the dissociation between recognition and categorization accuracy, and the comparable prototypicality gradients for memory-impaired patients and controls in categorization (see Nosofsky & Zaki, 1998, Figures 1 and 2). Variants of the GCM were used to successfully model the results of other studies of implicit category learning that have compared normal controls and memory-impaired patients.[7]

Analysis of the dissociation. Appreciating how the GCM works in producing the dissociation between categorization and recognition is useful. The first thing to note is that the model exploits a difference in how these two tasks were implemented in Knowlton and Squire (1993)—namely, that different items were used in the categorization and recognition tasks. Specifically, 40 different but similar dot patterns served as training items in categorization, whereas 5 dissimilar dot patterns were the memory items in recognition. This difference in materials allows the GCM model to readily capture the dissociation. Table 1 provides an illustration of how this works. The top half of the table illustrates the categorization task; the bottom half illustrates the recognition task. To keep things simple, for each task, only five training exemplars are represented (shown on the left side of Table 1), and each one is represented in terms of binary values on five dimensions, designated as D1 through D5 in the table. (As examples for dot patterns, one dimension might be whether there are many or few dots in the upper-left quadrant; for artificial animals, one dimension might be whether the creature has stripes or not.) The two test items for each task (shown on the right side of the table) are represented in the same way. The structure of the training and test items is intended to capture the structure of the items used by Knowlton and Squire.

For the categorization task (top half of Table 1), the exemplar representations required during training are highly similar to one another. Because of this intraset similarity, a positive test item (a category member) will be highly similar to most of the stored training exemplars, and it will have a high summed similarity: In contrast, a negative test item (a nonmember) may be highly dissimilar to the stored exemplars, and it will have a low summed similarity. The consequence is that it does not take much memory discrimination to tell members from nonmembers in categorization. The bottom half of Table 1 illustrates the recognition task. Now the training exemplars are less similar to one another, and the positive test items (old items) show only a moderate summed similarity to the stored exemplars, whereas the negative test items (new items) show a somewhat smaller summed similarity to the stored exemplars. The consequence is that it is relatively difficult to discriminate old items from new ones.

The upshot is that the difference between (1) summed similarity for positive test items and stored training exemplars and (2) summed similarity for negative test items and stored training exemplars is greater for categorization than for recognition. Consequently, if the outcome of a similarity comparison is diminished by decreasing sensitivity, as it presumably is in memory-impaired patients, then categorization may be unaffected, but recognition

Table 1
GCM Analysis of Memory Comparisons

	Training					Test				
	D1	D2	D3	D4	D5	D1	D2	D3	D4	D5
Categorization										
1.	1	1	1	1	0	1	1	1	1	1 (positive: member)
2.	1	1	1	0	1	0	0	0	0	1 (negative: nonmember)
3.	1	1	0	1	1					
4.	1	0	1	1	1					
5.	0	1	1	1	1					
Recognition										
1.	1	1	1	1	0	1	1	1	1	0 (positive: old)
2.	1	1	0	0	1	0	0	0	1	1 (negative: new)
3.	1	0	1	1	0					
4.	0	1	0	0	0					
5.	0	0	1	0	1					

should clearly suffer. This is how GCM predicts the dissociation. This analysis has not been addressed in previous attempts to rebut Nosofsky and Zaki (1998; see, e.g., Squire & Knowlton, 2000).[8]

A number of comments about the analysis above are in order. First, the differences in representations between categorization and recognition and the resulting similarity differences are sufficiently great so that the Nosofsky and Zaki (1998) model can capture the dissociation without recourse to any subtle interplay between free parameters in GCM. Indeed, the representational differences between categorization and recognition are such that other models of categorization based on explicit memory can also predict the critical dissociation. As an example, consider the SUSTAIN model proposed by Love, Medin, and Gureckis (2004). Unlike the exemplar-based GCM, SUSTAIN assumes that to the extent that training items are similar, participants will group them together into a single cluster, whereas dissimilar items will require their own clusters (i.e., they are treated like exemplars). When these ideas are applied to the training items represented on the left side of Table 1, it appears that only one cluster might be needed to represent the categorization items (the cluster characterized by the dimension values 11111), whereas three to four clusters might be needed to represent the recognition items. Assuming that patients with memory impairment are less likely to form a new cluster than are controls (which is captured by a single parameter in SUSTAIN), the model predicts that the patients will be impaired on recognition, but not on categorization (see Love & Gureckis, 2007).

Other existent mathematical models of category learning might also be comfortable with the dissociation. The general point is that the structure of the materials used in some relevant experiments are such that the results can be explained by a number of category learning models in a straightforward way. This criticism of key experiments on implicit category learning seems well founded.

Second, although the use of different materials for categorization and recognition leads to an interpretation problem, we note that there are some reasons for this design decision. One wants to use highly similar instances for the categorization task so as to mimic the structure of natural categories. However, employing the same instances in recognition could lead to problems. If the new items are similar to the training instances, the recognition task will be too difficult for patients, whereas if the new items are dissimilar to the training items, all participants may perform the recognition task by using a categorization strategy (see Bozoki et al., 2006). So the use of different materials in Knowlton and Squire (1993) at least ensured that the recognition task was no more difficult than the categorization task.

What about other prototype extraction experiments that have been reviewed? Either they too have used different materials for categorization and recognition (see, e.g., Squire & Knowlton, 1995), or they have used the same materials but have employed a memory task that did not permit direct comparison with categorization (e.g., Bozoki et al., 2006; Reed et al., 1999).

Clearly, we are in need of further neuropsychological experiments that use the same materials for categorization and recognition and that employ comparable tasks. If such experiments are performed and continue to produce a dissociation between categorization and recognition, then the case for implicit category learning will be strengthened, although it may still be possible for the GCM to account for the dissociation via parameters other than memory discrimination. In the version of GCM presented in Nosofsky and Zaki (1998), different parameters are used for the criteria for categorization and recognition, and these two parameters can take on different values for memory-impaired patients and controls. Variations in these parameters (and still others) could possibly yield the dissociation of interest. What is needed, then, are neuropsychological experiments that take into consideration the main component processes that are posited in current models of category learning. A similar argument is made in Love and Gureckis (2007).

Role of neuroimaging data. There is another source of evidence that can be brought to bear on the question of whether the results from prototype extraction and recognition tasks reflect the same underlying mechanisms—neuroimaging evidence. Using functional magnetic resonance imaging (fMRI), we can ask whether prototype extraction and recognition tasks recruit different neural networks, even in normal controls (this kind of research has not yet been done with patients).

Reber, Stark, and Squire (1998) performed exactly this kind of experiment, using the same tasks that were employed in Knowlton and Squire (1993) (see Figure 1A), but with all of the participants being normal. In prototype extraction, participants first were presented 40 different distortions of a prototype pattern, and then categorized 70-plus novel patterns while being imaged by fMRI. In recognition, participants first were presented five patterns (different from those in prototype extraction), and then they made old–new judgments while being imaged by fMRI. The behavioral results indicated that, if anything, categorization again was the more difficult task (categorization accuracy was only 58%, whereas the hit rate in recognition was 81%). Of greater importance are the imaging results. In categorization, the fMRI contrast of interest was the difference in activation between a category member and a nonmember (member minus nonmember); in recognition, the contrast of interest was the difference in activation between an old item and a new one (old minus new). These contrasts produced striking differences between the two tasks. The recognition task resulted in numerous activations, including several areas in the prefrontal cortex and the MTL. These are exactly the areas that have repeatedly been implicated in neuroimaging studies of explicit memory (for recent reviews, see Squire, Clark, & Bayley, 2004; Wagner, Bunge, & Badre, 2004). But none of these areas was activated in categorization. Even more dramatically, an area in the posterior occipital cortex known to be involved in visual processing (Brodmann Area 17/18) was activated in recognition but deactivated in categorization. At face value, these results provide evidence that different memory systems are involved in prototype extraction and recognition.

For our purposes, however, this study has a number of critical limitations. For one, the difference in materials used in the two tasks leaves open the possibility that the greater activations in recognition than in categorization may reflect the greater demands on memory discrimination in the recognition task. (It is less clear, however, how the difference in material could explain the finding that there were some different activations in the two tasks.) A second problem with the study is that the critical contrasts for categorization and recognition may not be strictly comparable; the difference between a category member and a nonmember is somewhat arbitrary in this task (see note 1) and may be different from the nonarbitrary difference between an old and a new item.

A more recent fMRI study by Reber et al. (2003) solved these two problems by using only prototype extraction but varying whether the task was performed with the usual incidental instructions (participants did not know a category was present at the beginning of training) or, instead, with intentional instructions (participants were told that there was a category during training). Because the items were identical in both conditions, and because the only fMRI contrast was between members and nonmembers at test, Reber et al. (2003) eliminated the two problems mentioned above. Moreover, in asking whether incidental and intentional instructions recruit different neural systems in categorization, Reber et al. (2003) asked a question that is even more germane to the issue of dual-category learning systems than is the question of categorization versus recognition. Intentional instructions should engage explicit category learning, which is mediated by structures in the MTL (particularly the hippocampus). To the extent that incidental instructions show a qualitatively different pattern of neural activation, there is evidence for a second system.[9]

Two different groups of participants were used. During training, the incidental group was told to point to the center of each dot pattern, whereas the intentional group was told that the patterns all came from the same category and that they were to learn it. Participants were imaged by fMRI during the test trials when both groups were instructed to discriminate category members from nonmembers. The behavioral results showed that the intentional instructions led to substantially better categorization performance than did the incidental instructions. Such an effect is hardly informative about whether there are separate explicit and implicit category-learning systems, since the difference in accuracy could merely reflect a difference in the amount of attention paid during training (a process that is included in virtually all single-system mathematical models of category learning). But the imaging results are more informative on the issue of multiple learning systems. In line with other findings on category learning (see, e.g., Koenig et al., 2004), intentional instructions led to increased activations in a number of brain areas, including prefrontal and parietal areas; in line with the Reber et al. (1998) study discussed above, incidental instructions led to a deactivation in the visual cortex. The fact that a difference in instructions determined whether the pattern was one of activation or deactivation suggests that the different instructions recruit qualitatively different categorization systems. (See Aizenstein et al., 2000, for comparable results with a different version of prototype extraction.)

In a particularly diagnostic contrast, Reber et al. (2003) focused on neural activity in two target regions: the hippocampus (a critical part of the MTL system) and the posterior occipital region that was deactivated in the incidental condition (as well as in Reber et al., 1998). Substantial research indicates that in comparison with an appropriate baseline, hippocampal activation is a marker of explicit memory, whereas deactivation of the extrastriate occipital cortex is a marker of implicit memory (as reflected in perceptual priming; see Buckner, 2000, for a relatively recent review). The results of this contrast are reproduced in Figure 3 and show a double dissociation: In the hippocampus, intentional instructions led to greater activity than did incidental instructions (incidental instructions produced no discernible activation here); in the occipital area, incidental instructions led to a greater deactivation than did intentional instructions. These results provide data that cannot easily be accounted for by any single-system model of category learning. To the extent to which these findings prove robust, they offset the challenge raised by the Nosofsky and Zaki (1998) analysis.

Although these imaging data are informative, by no means are they the last word on the matter. There are no direct comparisons of the hippocampal activation or the occipital deactivation across the two tasks; rather, the activations and deactivations are compared only with a baseline. There are also some anomalies in the data. For one, in Reber et al. (2003), the intentional condition shows some deactivation of the posterior occipital site, whereas in the earlier Reber et al. (1998) study, the recognition task con-

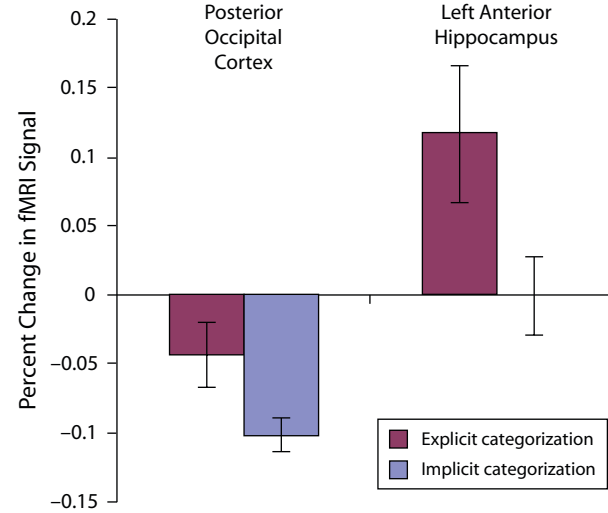

Figure 3. Reber, Gitelman, Parrish, and Mesulam (2003) experiment: The fMRI difference between members and nonmembers in the intentional (referred to as "explicit") and unintentional (referred to as "implicit") conditions. From "Dissociating Explicit and Implicit Category Knowledge With fMRI," by P. J. Reber, D. R. Gitelman, T. B. Parrish, and M. M. Mesulam, *Journal of Cognitive Neuroscience*, *15*, p. 578. Copyright 2003 by MIT Press. Adapted with permission.

dition (which should be similar to intentional categorization) shows a significant activation of the same area. Another issue concerns hippocampal activation. Reber et al. (2003) found no such activation in their incidental condition, but we recently have obtained hippocampal activation in a comparable task (Koenig et al., 2007). According to arguments made earlier in this article, there should be such activation even with incidental instructions, because the present participants were normal controls, and such participants seem to employ explicit memory in the prototype extraction task. The upshot of these two points is that the apparent double dissociation displayed in Figure 3 may be more a quantitative than a qualitative difference; there may be some hippocampal activation with incidental instructions (but less than with intentional instructions), and some occipital deactivation with intentional ones (but less than with incidental instructions). Even a quantitative difference, however, challenges the Nosofsky and Zaki (1998) model-based claim that a single system can explain categorization.

There is a superficial sense in which imaging data cannot be explained by some of the mathematical models that we have considered. As currently formulated, some of the models make no claims about the neural bases of cognitive functions and, hence, have nothing to say about neural data. But the imaging data above are incompatible with existent single-system models in a deeper sense. There is so much evidence linking explicit memory tasks to the hippocampus that any cognitive model that claims that category learning is in essence an explicit memory task seems obligated to predict hippocampal activation during category learning. (Indeed, Nosofsky and Zaki have consistently accepted the link between explicit memory and hippocampal function when discussing studies with MTL patients; see, e.g., Nosofsky & Zaki, 1998, 1999.) There are some correspondences between cognitive processes and brain areas that are so well established that neither a personal disavowal of interest in the brain nor a principled skepticism about the limits of reductionism seems an acceptable reason to ignore neural evidence.

Summary of criticisms and counterarguments. We have considered five criticisms of the evidence for the existence of implicit category learning. The first two of these argued that the results obtained in prototype extraction can be attributed to learning during test trials via working memory mechanisms that are relatively intact in memory-impaired patients (Palmeri & Flanery, 1999; Zaki & Nosofsky, 2001). These two criticisms correctly point out a problem with the standard prototype extraction paradigm—namely, that the test trials are sufficiently numerous and structured so as to permit learning. But although such learning clearly occurs, the recent results of Bozoki et al. (2006) indicate that such learning is not always sufficient to account for the overall performance of memory-impaired patients in prototype extraction; learning during training also makes a contribution, and this learning presumably is implicit.

The second criticism also argued that the results obtained in at least one prototype extraction task (Reed et al., 1999) can be attributed to participants' explicitly remembering one or two features from the training phase. But the evidence supplied was too indirect to convincingly overthrow the behavioral findings that show that neither patients nor controls restricted their categorization decisions to only one or two features.

The third criticism argued that contrary to the findings in individual experimental reports, meta-analyses reveal that memory-impaired patients are somewhat impaired on tasks used to study implicit category learning (Zaki, 2004). These meta-analyses lack the data to make a strong case that memory-impaired patients are impaired on prototype extraction in particular. More importantly, the single-system account that Zaki applies to some of these experiments seems to lack plausibility.

The fourth criticism holds that the prototype extraction categorization is simply a less difficult task than the recognition tasks with which it has been compared (Zaki et al., 2003). This claim is at odds with the findings that prototype extraction actually is easier than recognition in most of the relevant studies, and the experiment that Zaki et al. offer in support of the criticism does not meet the boundary conditions of the kind of implicit category learning considered in this article.

The most telling criticism of research on implicit category learning is the Nosofsky and Zaki (1998) demonstration that a single-system explicit memory model can readily account for normal and patient data on both prototype extraction and recognition; this comes about because the categorization task is less taxing than the recognition task of a memory discrimination process that is a likely subcomponent of both tasks. This criticism highlights the need for detailed computational accounts of tasks used in neuropsychology and makes clear the problems that arise when the unit of analysis is a task, rather than a component process. But as useful as the criticism is, the single-system model of Nosofsky and Zaki (1998) seems at odds with recent neuroimaging evidence, which reveals different patterns of activations and deactivations for categorization when it is based on explicit versus implicit processes.

All things considered, the evidence favors the existence of an implicit category-learning system.

THE NATURE OF IMPLICIT CATEGORY LEARNING

Given that there is substantial evidence for an implicit category-learning system, what is the nature of the processes involved? The neuroimaging findings suggest an answer. The deactivations observed in the posterior occipital cortex are similar to those routinely reported in imaging studies of perceptual priming (e.g., Buckner, 2000). In the latter studies, the deactivation is attributed to increased efficiency in processing the perceptual features of the stimulus. Thus, when one reads a word faster the second time it is presented, presumably the speedup reflects faster processing of the features that comprise the letters of the word. The same kind of process may be in play in prototype extraction. Because the training items are highly similar to one another, they must have numer-

ous features in common. These common features, by definition, occur frequently during training, and consequently, the participants in these experiments become increasingly proficient in processing them. In turn, the participants may experience a feeling/sensation of greater perceptual fluency with items that contain many common features. Then—and this is the critical point—during the test phase, participants can use this feeling of perceptual fluency as a basis of categorization: the greater the fluency, the more likely it is that the item is a member of the category (see also Squire & Knowlton, 2000).

This perceptual-fluency hypothesis seems compatible with all the critical data. First, it explains why memory-impaired patients are relatively intact on categorization but are impaired on recognition. Specifically, (1) judgments made on the basis of fluency may be sufficient for single-category categorization, but they are not a reliable indicator for recognition (new items can be composed of the same features as old ones, but in different combinations), and (2) patients with memory impairment are capable of increasing their processing efficiency of frequently occurring features (they are intact in perceptual priming), so they have the critical ingredient of the perceptual fluency process.

Second, the perceptual fluency hypothesis directly predicts prototypicality gradients. The more prototypical a test item, the more common features it has, the more it will be accompanied by a greater feeling of perceptual fluency, and the more likely it is to be categorized as a member. Moreover, because this account hinges only on the perceptual fluency mechanism, prototypicality gradients should be comparable for normal controls and patients with memory impairments.

Third, a perceptual fluency process is consistent with increasing deactivation in brain areas associated with perceptual processing—the less processing needed, the less the activation.

Fourth, because perceptual fluency is based on relatively low-level perceptual processes that people are not aware of, the hypothesis is consistent with the observation that, typically, people cannot articulate implicit knowledge (see, e.g., Roediger & McDermott, 1993).

Perceptual fluency is a step up from perceptual priming; it assumes that one gets feedback about how effortful one's processing is and then uses this feedback to make categorical judgments. The process is sufficient to support the categorization of novel items, which is an important criterion for claiming that a category has been learned. What about other criteria for category knowledge? One is that what has been learned can be used with items presented in a different modality or format. Could the implicit knowledge obtained in prototype extraction with artificial animals be used if the animals' features were instantiated differently (e.g., different sized spots were used on different animals)? To the extent perceptual fluency operates on truly low-level perceptual properties, the mechanism may not meet this condition. Another criterion for category knowledge is that it support inductive inferences (see, e.g., Smith, 1995). If one has acquired a category of artificial animals via perceptual fluency and then learns that some of the instances have an additional feature, will one generalize this feature to other category members? Attempting to answer questions such as these will require the development of new paradigms for studying implicit category learning.

SUMMARY

All things considered, there is indeed evidence for implicit category learning. The seminal neuropsychological studies of Knowlton and Squire (1993), as well as later follow-ups, have provided suggestive evidence for such a system. But critiques from cognitive psychology and mathematical modeling have raised important challenges to the neuropsychological findings. Some of the critiques have focused on the role of working memory in categorization tasks, but the challenges raised have been at least partially offset by the recent Bozoki et al. (2006) results. Perhaps the most important critique of the evidence for implicit category learning was contained in Nosofsky and Zaki's (1998) analysis of the shared-component processes in categorization and recognition; their discussion reminds us that the unit of analysis in neuropsychology, as in mainstream cognitive psychology, must be the component processes, not the task. But recent imaging evidence (e.g., Reber et al., 2003) provides evidence that offsets the Nosofsky and Zaki (1998) critique. This evidence also suggests that the critical mechanism underlying implicit category learning is perceptual fluency. What becomes of interest, then, is whether this mechanism can yield truly categorical knowledge.

AUTHOR NOTE

Preparation of this article, as well as the execution of some of the studies cited in it, was supported in part by U.S. Public Health Service Grants AG17586, AG15116, and NS35867. The author is indebted to his colleagues, Murray Grossman, Brad Love, and Janet Metcalfe, for their helpful comments on an earlier version of this article. The author also acknowledges the insightful and thorough comments of the three external reviewers of this article (Rob Nosofsky, Russ Poldrack, and one anonymous reviewer). Address correspondence to E. E. Smith, Department of Psychology, Columbia University, 1190 Amsterdam Ave., MC 5501, New York, NY 10027 (e-mail: eesmith@psych.columbia.edu).

Note—This article was accepted by the previous editorial team, when John Jonides was Editor.

REFERENCES

AIZENSTEIN, H. J., MACDONALD, A. W., STENGER, V. A., NEBES, R. D., LARSON, J. K., URSU, S., ET AL. (2000). Complementary category learning systems identified using event-related functional MRI. *Journal of Cognitive Neuroscience*, **12**, 977-987.

ASHBY, F. G., & ELL, S. W. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences*, **5**, 204-210.

ASHBY, F. G., & MADDOX, W. T. (2005). Human category learning. *Annual Review of Psychology*, **56**, 149-178.

ASHBY, F. G., & O'BRIEN, J. B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Sciences*, **9**, 83-89.

BOZOKI, A., GROSSMAN, M., & SMITH, E. E. (2006). Can patients with Alzheimer's disease learn a category implicitly? *Neuropsychologia*, **44**, 816-827.

BUCKNER, R. L. (2000). Neuroimaging of memory. In M. Gazzaniga (Ed.), *The new cognitive neurosciences* (2nd ed., pp. 1013-1022). Cambridge, MA: MIT Press.

FOERDE, K., KNOWLTON, B. J., & POLDRACK, R. A. (2006). Modulation of competing memory systems by distraction. *Proceedings of the National Academy of Sciences*, **103**, 11778-11783.

KERI, S. (2003). The cognitive neuroscience of category learning. *Brain Research Reviews*, **43**, 85-109.

KERI, S., KALMAN, J., KELEMEN, O., BENEDEK, G., & JANKA, Z. (2001). Are Alzheimer's disease patients able to learn visual prototypes? *Neuropsychologia*, **39**, 1218-1223.

KERI, S., KALMAN, J., RAPCSAK, S. Z., ANTAL, A., BENEDEK, G., & JANKA, Z. (1999). Classification learning in Alzheimer's disease. *Brain*, **122**, 1063-1068.

KNOWLTON, B. J. (1999). What can neuropsychology tell us about category learning? *Trends in Cognitive Sciences*, **3**, 123-124.

KNOWLTON, B. J., MANGELS, J. A., & SQUIRE, L. R. (1996). A neostriatal habit learning system in humans. *Science*, **273**, 1399-1402.

KNOWLTON, B. J., & SQUIRE, L. R. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science*, **262**, 1747-1749.

KOENIG, P., SMITH, E. E., GLOSSER, G., DEVITA, C., MOORE, P., MCMILLAN, C., ET AL. (2004). The neural basis for novel semantic categorization. *NeuroImage*, **24**, 369-383.

KOENIG, P., SMITH, E. E., TROIANI, V., ANTANI, S., MCCAWLEY, G., MOORE, P., ET AL. (2007). *Collaborating implicit and explicit memory mechanisms: Evidence from Alzheimer's disease and fMRI.* Manuscript submitted for publication.

KOLODNY, J. A. (1994). Memory processes in classification learning: An investigation of amnesic performance in categorization of dot patterns and artistic styles. *Psychological Science*, **5**, 164-169.

LOVE, B. C., & GURECKIS, T. N. (2007). Models in search of a brain. *Cognitive, Affective, & Behavioral Neuroscience*, **70**, 90-108.

LOVE, B. C., MEDIN, D. L., & GURECKIS, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, **111**, 309-332.

MAY, C. P., HASHER, L., & FOONG, N. (2005). Implicit memory, age, and time of day: Paradoxical priming effects. *Psychological Science*, **16**, 96-100.

NOSOFSKY, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 700-708.

NOSOFSKY, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception & Performance*, **17**, 3-27.

NOSOFSKY, R. M., & ZAKI, S. R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. *Psychological Science*, **9**, 247-255.

NOSOFSKY, R. M., & ZAKI, S. R. (1999). Math modeling, neuropsychology, and category learning: Response to B. Knowlton (1999). *Trends in Cognitive Sciences*, **3**, 125-126.

PALMERI, T. J., & FLANERY, M. A. (1999). Learning about categories in the absence of training: Profound amnesia and the relationship between perceptual categorization and recognition memory. *Psychological Science*, **10**, 526-530.

POLDRACK, R. A., CLARK, J., PARÉ-BLAGOEV, E. J., SHOHAMY, D., CRESO MOYANO, J., MYERS, C., & GLUCK, M. A. (2001). Interactive memory systems in the human brain. *Nature*, **414**, 546-550.

POSNER, M. I., & KEELE, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, **77**, 353-363.

REBER, P. J., GITELMAN, D. R., PARRISH, T. B., & MESULAM, M. M. (2003). Dissociating explicit and implicit category knowledge with fMRI. *Journal of Cognitive Neuroscience*, **15**, 574-583.

REBER, P. J., STARK, C. E. L., & SQUIRE, L. R. (1998). Contrasting cortical activity associated with category memory and recognition memory. *Learning & Memory*, **5**, 420-428.

REED, J. M., SQUIRE, L. R., PATALANO, A. L., SMITH, E. E., & JONIDES, J. J. (1999). Learning about categories that are defined by object-like stimuli despite impaired declarative memory. *Behavioral Neuroscience*, **113**, 411-419.

ROEDIGER, H. L., III, & MCDERMOTT, K. B. (1993). Implicit memory in normal human subjects. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology* (Vol. 8, pp. 63-131). New York: Elsevier.

SCHACTER, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **13**, 501-518.

SEGER, C. A., & CINCOTTA, C. M. (2002). Striatal activity in concept learning. *Cognitive, Affective, & Behavioral Neuroscience*, **2**, 149-161.

SHOHAMY, D., MYERS, C. E., GROSSMAN, S., SAGE, J., GLUCK, M. A., & POLDRACK, R. A. (2004). Cortico-striatal contributions to feedback-based learning: Converging data from neuroimaging and neuropsychology. *Brain*, **127**, 851-859.

SMITH, E. E. (1995). Concepts and categorization. In E. E. Smith & D. Osherson (Eds.), *Invitation to cognitive science: Vol. 3. Thinking* (2nd ed., pp. 3-33). Cambridge, MA: MIT Press.

SMITH, E. E., PATALANO, A., & JONIDES, J. (1998). Alternative mechanisms of categorization. *Cognition*, **65**, 167-196.

SQUIRE, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning & Memory*, **82**, 171-177.

SQUIRE, L. R., CLARK, R. E., & BAYLEY, P. J. (2004). Medial temporal lobe function and memory. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (3rd ed., pp. 691-708). Cambridge, MA: MIT Press.

SQUIRE, L. R., & KNOWLTON, B. J. (1995). Learning about categories in the absence of memory. *Proceedings of the National Academy of Sciences*, **92**, 12470-12474.

SQUIRE, L. R., & KNOWLTON, B. J. (2000). The medial temporal lobe, the hippocampus, and the memory systems of the brain. In M. Gazzaniga (Ed.), *The new cognitive neurosciences* (2nd ed., pp. 756-776). Cambridge, MA: MIT Press.

WAGNER, A. D., BUNGE, S. A., & BADRE, D. (2004). Cognitive control, semantic memory, and priming: Contributions from prefrontal cortex. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (3rd ed., pp. 709-725). Cambridge, MA: MIT Press.

ZAKI, S. R. (2004). Is categorization performance really intact in amnesia? A meta-analysis. *Psychonomic Bulletin & Review*, **11**, 1048-1054.

ZAKI, S. R., & NOSOFSKY, R. M. (2001). A single-system interpretation of dissociations between recognition and categorization in a task involving object-like stimuli. *Cognitive, Affective, & Behavioral Neuroscience*, **1**, 344-359.

ZAKI, S. R., NOSOFSKY, R. M., JESSUP, N. M., & UNVERZAGT, F. W. (2003). Categorization and recognition performance of a memory-impaired group: Evidence for single-system models. *Journal of the International Neuropsychological Society*, **9**, 394-406.

NOTES

1. The fact that the categorization task is named "prototype extraction" does not mean that that is the cognitive mechanism involved. The notion of accuracy in the categorization task is somewhat arbitrary, because no feedback was given. A test item was considered a category member if it was the prototype or a systematic distortion of it, and a nonmember otherwise.

2. Strictly speaking, single-system theorists like Nosofsky and Zaki (1998) would not refer to the memory system involved as "explicit," because they see no need to draw the explicit–implicit distinction in the first place. I use the qualifier "explicit" when referring to their claims mainly to keep things clear.

3. Zaki and Nosofsky (2001) partially replicated these results with artificial animals. Specifically, they tested 37 normal participants in a no-training condition and found that 23 of them showed evidence for category learning during test, whereas the remaining 14 participants performed at chance. It appears that if Zaki and Nosofsky had averaged the data from all their participants, they would have produced a shallow prototypicality gradient that is similar to that found by Palmeri and Flanery (1999).

4. Note that both Knowlton and Squire (1993) and Reed et al. (1999) did try to determine whether their participants could learn the category solely from the test trials. In both studies, a separate group of controls was instructed to imagine that a set of training trials had been presented and were then given the standard test trials. Neither study found any evidence for category learning in the absence of training trials. Presumably, these failures were due to the fact that the imagine instructions did not convince the participants that they could learn the category, in contrast to the Palmeri and Flanery (1999) cover story about subliminal presentation of training items.

5. A qualification about the Bozoki et al. (2006) study is in order. In addition to the difference between test performance with and with-

out training, Bozoki et al. used a second measure of implicit category learning—namely, performance on the first 10 test trials. Presumably, working memory mechanisms should have had relatively little time to operate during these early trials. The controls performed significantly better than the patients on this measure. But note that this second measure of implicit category learning was based on roughly one eighth of the data that went into the first measure (the difference between performance with and without training); for this reason, Bozoki et al. favored the first measure.

6. This alternative account of Zaki (2004) implies that we cannot expect to find true dissociations between performance in prototype extraction paradigms and performance in standard tests of explicit memory, such as recognition memory. Interestingly, the same point may be true for would-be dissociations between probabilistic categorization and tests of explicit memory, since recent findings indicate that normal participants use some explicit memory in the most widely used paradigm for studying probabilistic categorization (Foerde et al., 2006).

7. This is the same general model that was used by Zaki and Nosofsky (2001) to argue that memory-impaired patients used only one or two features in the Reed et al. (1999) study (see the subsection on *Do Participants Explicitly Remember Only a Few Features?*). The criticisms offered against that instantiation of the model do not apply here; for example, in the current instantiation, empirical estimates of critical parameters were obtained, thereby constraining parameters, and one version of the model contained only three authentically free parameters.

8. The analysis captures most of the story for memory-impaired patients but leaves something out for normal controls. Because controls have substantial memory discrimination, they will find some exact matches during recognition (but not during categorization). And an additional feature of the model—the parameter for memory discrimination is transformed by a power function— guarantees that exact matches contribute disproportionally to the summed similarities. For this reason, recognition may be easier than categorization for controls.

9. Reber et al. (2003) seem to be assuming that when items are processed under minimalist incidental instructions, the only memory formed is implicit. The same assumption has been made in all prototype extraction studies but has never been systematically tested.

(Manuscript received August 8, 2006;
revision accepted for publication February 2, 2007.)