Perception & Psychophysics
2008, 70 (6), 983-999
doi: 10.3758/PP.70.6.983

A role for the perceptual representation memory system in category learning

Michael B. Casale and F. Gregory Ashby
University of California, Santa Barbara, California

There is growing evidence that working memory, episodic/semantic memory, and procedural memory all play important roles in at least some types of category learning. Little is known, however, about the role of the perceptual representation memory system (PRS). Two experiments are reported that provide evidence that under certain conditions, the PRS, by itself, is sufficient to mediate category learning. Both experiments compared performance in (A, not A) and (A, B) prototype distortion category-learning tasks, in which category exemplars are created by randomly distorting one category prototype in the (A, not A) conditions or two prototypes in the (A, B) conditions. Results showed that (A, not A) performance was more sensitive to prototype similarity and less affected by the removal of feedback than (A, B) performance. These results support the hypothesis that (A, not A) performance was mediated by the PRS, but that (A, B) performance recruited other memory systems.

There is growing evidence and theoretical speculation that all major memory systems contribute to category learning (Ashby & O'Brien, 2005). For example, empirical evidence suggests that at least some types of category learning are mediated by working memory (DeCaro, Thomas, & Beilock, 2008; Maddox, Ashby, Ing, & Pickering, 2004; Waldron & Ashby, 2001; Zeithamova & Maddox, 2006), episodic/semantic memory (Hopkins, Myers, Shohamy, Grossman, & Gluck, 2004; Knowlton, Squire, & Gluck, 1994; Kolodny, 1994; Zaki, Nosofsky, Jessup, & Unversagt, 2003), or procedural memory (Ashby, Ell, & Waldron, 2003; Maddox, Bohil, & Ing, 2004). The one conspicuously absent system in this list is the perceptual representation memory system, or PRS (Schacter, 1990). There is empirical evidence that the PRS is active during perceptual categorization (Aizenstein et al., 2000; Reber, Stark, & Squire, 1998a, 1998b), but this evidence is from functional neuroimaging, not behavioral data. More specifically, we know of no evidence that the PRS mediates category learning performance. The goal of this article is to report such evidence.

The PRS has been described as the memory system that mediates "improvement in identifying or processing a stimulus as the result of its having been observed previously" (Gazzaniga, Ivry, & Mangun, 2002). The type of learning mediated by the PRS is often referred to as *repetition priming*. The PRS is an implicit system that can operate without conscious awareness (Schacter, 1990), and behavioral effects of the PRS can be observed after only a single stimulus repetition (e.g., Wiggs & Martin, 1998). Furthermore, the duration of these effects are long lasting. For example, Cave (1997) demonstrated that behavioral effects of the PRS can be seen as long as 48 weeks after a single stimulus (i.e., picture) presentation. Another finding relevant to the categorization literature is that PRS effects can be induced when the two stimuli are different but perceptually similar (e.g., Biederman & Cooper, 1992; Cooper, Schacter, Ballesteros, & Moore, 1992; Seamon et al., 1997).

These results suggest that the PRS should be operating in most perceptual categorization experiments. This should be true for any experiment in which the presentation of category exemplars is repeated, or in which a category contains multiple exemplars that have high perceptual similarity. As mentioned above, functional neuroimaging data support this prediction (Aizenstein et al., 2000; Reber et al., 1998a, 1998b). What is not so clear is whether the PRS by itself is ever sufficient to mediate the categorization process; in other words, is it ever possible for participants to use the PRS and no other major memory system when learning which response is associated with each stimulus in a categorization task?

Consider an experiment with two categories denoted A and B. Suppose each category contains either exemplars that are each presented equally often during the experiment, or exemplars in which within-category similarity is roughly equal in the two categories. In either case, we could expect the PRS to be equally active on both A and B trials. In its original description, Schacter (1990) argued that the PRS did not "represent elaborative information that links an event to pre-existing knowledge" (p. 553). Instead, he proposed that the PRS could provide "a basis for a feeling of familiarity" (p. 553). Thus, in either of these experiments, after just a few trials, the PRS could signal to the participant that the stimulus was familiar; but this is all it could signal. For example, the PRS is not thought to encode an explicit memory of the category

F. G. Ashby, ashby@psych.ucsb.edu

prototype or of any other previously seen category exemplar. Presentation of the prototype should quickly elicit a feeling of familiarity from the PRS, but presentation of a stimulus that is perceptually distinct from the prototype would not elicit a memory of the prototype from the PRS. This would require some other memory system. Therefore, other memory systems would be required to work in collaboration with the PRS to signal the participant which response to make in these experiments.

Now consider another experiment with two categories labeled A and B. Suppose now that Category A is the same as before—that is, it either contains a small number of exemplars that are repeated, or it is highly coherent and within-category similarity is high. In contrast, suppose Category B contains many exemplars that are never repeated and that within-category similarity is so low that every pair of exemplars is highly distinct. In this case, the PRS will be active on Category A trials, but not on Category B trials. Therefore, the participant should quickly develop a feeling that stimuli on A trials seem familiar, but stimuli on B trials do not. In this case, participants could feasibly adopt a decision rule of the following type: "If the stimulus feels familiar, I'll respond 'A'; if it feels unfamiliar, I'll respond 'B.'" This decision rule depends only on the PRS.

Most categorization experiments are of the former type, in which case the PRS is predicted to be equally active on A and B trials. We predict that it should be difficult to find evidence from such experiments that the PRS is mediating the learning of category responses. Instead, we hypothesize that evidence for a role of the PRS in category learning should focus on experiments of the latter type—that is, on experiments in which the PRS should be much more active on A trials than on B trials (or vice versa).

It turns out that there is a popular categorization paradigm that commonly includes both types of task. In prototype distortion category-learning tasks, the category exemplars are created by randomly distorting a single category prototype. The most widely known example uses a constellation of dots (often 7 or 9) as the category prototype, and the other category members are created by randomly perturbing the spatial location of each dot. These random dot stimuli and categories have been used in dozens of studies (e.g., Homa, Rhoads, & Chambliss, 1979; Homa, Sterling, & Trepel, 1981; Posner & Keele, 1968, 1970; Shin & Nosofsky, 1992; Smith & Minda, 2002).

Two different types of prototype distortion tasks are commonly used—(A, B) and (A, not A). In an (A, B) task, participants are presented a series of exemplars that are each from some Category A, or from a contrasting Category B. The task of the participant is to respond with the correct category label on each trial (i.e., "A" or "B"). An important feature of (A, B) tasks is that the stimuli associated with both responses each have a coherent structure—that is, they each have a central prototypical member around which the other category members cluster. Thus, within-category similarity is equally high in both categories in (A, B) prototype distortion tasks. In an (A, not A) task, on the other hand, there is a single central Category A and participants are presented with stimuli that are either exemplars from Category A or random patterns that do not belong to Category A. The participant's task is to respond "Yes" or "No" depending on whether the presented stimulus was or was not a member of Category A. In an (A, not A) task, the Category A members have a coherent structure since they are created from a single prototype, but the stimuli associated with the "not A" (or "No") response do not. Typically, the two stimuli in every pair of "not A" category members are visually distinct. Historically, prototype distortion tasks have been run in both (A, B) and (A, not A) forms, although (A, not A) tasks are most common.

We can summarize our arguments so far by stating our main hypothesis—the PRS should facilitate performance in (A, not A) prototype distortion tasks, especially when the level of distortion is low (and within-category similarity is high); but the PRS by itself cannot mediate performance improvements in (A, B) prototype distortion tasks. To our knowledge, this hypothesis was first proposed by Ashby and Casale (2003; see also Reber & Squire, 1999), and has received no behavioral tests. The goal of this article is to test this hypothesis.

There is some indirect support for this PRS hypothesis. First, a variety of neuropsychological patient groups that are known to have widespread category-learning deficits show apparently normal (A, not A) prototype distortion learning. This includes patients with Parkinson's disease (Reber & Squire, 1999), schizophrenia (Kéri, Kelemen, Benedek, & Janka, 2001), or Alzheimer's disease (Sinha, 1999; although see Kéri et al., 1999). Second, several studies have reported normal (A, not A) prototype distortion learning in patients with amnesia (Knowlton & Squire, 1993; Squire & Knowlton, 1995), but impaired performance in (A, B) tasks (Zaki et al., 2003).

Third, neuroimaging studies of (A, not A) prototype distortion tasks have all reported categorization-related changes within occipital cortex (Aizenstein et al., 2000; Reber et al., 1998a, 1998b). In the only known neuroimaging study of the (A, B) prototype distortion task, Seger et al. (2000) also reported categorization-related activation in occipital cortex, but they also found significant learning-related changes in prefrontal and parietal cortices. Occipital cortex deactivations are often seen in tasks that depend on the PRS (e.g., Wiggs & Martin, 1998), and these neuroimaging results have prompted proposals that the PRS is active in prototype distortion tasks (Reber & Squire, 1999). On the other hand, such deactivations are typically not correlated with behavioral measures (Schacter, Wig, & Stevens, 2007), so the neuroimaging data do not speak to the question of whether the PRS is mediating prototype distortion *learning*.

What would constitute evidence that the PRS is mediating learning in (A, not A) tasks? One empirical signature of the PRS is that it should be more sensitive to distortion than most declarative strategies. Increasing distortion decreases similarity to the prototype, and there is evidence that the PRS is highly sensitive to reductions in similarity. For example, PRS activation is reduced (but not eliminated) if the second presentation of a word is in a different font from the first presentation (e.g., Jacoby & Hayman, 1987; Roediger & Blaxton, 1987). Similar reductions in PRS activation are also seen if the second presentation of an object is in

a different color or seen from a different viewpoint from the first presentation (Biederman & Gerhardstein, 1993; Cave, Bost, & Cobb, 1996), or if it is a different token of the object presented first (Cave et al., 1996). For example, Koutstaal et al. (2001) used fMRI to compare PRS activation (i.e., repetition suppression; see the General Discussion for more details) when the second presentation of an object was identical to the first, as opposed to when the second presentation was a different token of the same object (e.g., a different umbrella from the one first presented). Koutstaal et al. (2001) reported reliable PRS activation when the second presentation was a different token, but the magnitude of this effect was about four times smaller than when the second presentation was identical to the first (e.g., Koutstaal et al., 2001, Figure 4, p. 193).

In each of these cases, the manipulated features were irrelevant to the task (e.g., font of the word, color of the object), and there is widespread evidence that changing a relevant feature reduces PRS activation much more than changing an irrelevant feature (e.g., Roediger & Srinivas, 1993; Wiggs & Martin, 1998). In prototype distortion tasks (including our experiments) all relevant features are distorted when creating the exemplars of Category A. Thus, the PRS literature suggests that Category A patterns that are dissimilar to the A prototype should only weakly activate the PRS. As a result, if the PRS is mediating performance, then the probability of responding "A" to a particular pattern should decrease quickly with the dissimilarity of that pattern to the A prototype.

In contrast, categorization strategies that depend on declarative memory will tend to predict that this probability decreases more slowly as dissimilarity increases. For example, suppose that participants notice that Category A exemplars are often characterized by some feature formed by a subset of the nine dots. One possibility might be a feature such as "the belt of Orion." Because such features depend only on the location of a subset of the dots (three in the case of the belt of Orion), they will be present in all low-distortion patterns, but also in some high-distortion patterns (e.g., those where the dots that are not critical to the feature are distorted more than the subset of dots that define the feature). Thus, participants looking for distinctive features of this type should respond "A" to many high-distortion patterns. In other words, any pattern that strongly activates the PRS should display a distinctive feature such as the belt of Orion, but some patterns displaying that feature will not activate the PRS. Similarly, Smith and Minda (2001) showed that an exemplar-based decision strategy, in which participants compare the stimulus to stored representations of previously seen exemplars, also predicts a relative insensitivity to prototype similarity. In particular, exemplar models predict that the probability of an "A" response should decrease slowly as the dissimilarity between the stimulus and prototype increases.

Experiment 1 tests these predictions in a 2 × 2 experiment, in which the two types of prototype distortion tasks [(A, not A) vs. (A, B)] are crossed with two levels of distortion (low vs. high). In each experimental condition, the critical dependent variable is the proportion of correct "A" responses as a function of the relative dissimilarity to the A prototype. The PRS hypothesis makes the strong prediction that these endorsement proportions should decrease (with relative dissimilarity to the prototype) more quickly in the (A, not A) condition than in the (A, B) condition. This is because the PRS should be especially sensitive to distortion, and because we predict that the PRS could facilitate (A, not A) performance more than (A, B) performance.

Note that the critical prediction relates to the *relative* dissimilarity of the stimulus to the A prototype. In the (A, not A) task, an A stimulus becomes more difficult to categorize as its distance to the A prototype increases (e.g., see Equation 2, below). In the (A, B) conditions, however, moving an A stimulus away from the A prototype increases difficulty only if the movement reduces the distance to the B prototype (see, e.g., Equation 1). Thus, our main focus will be on the probability of correctly responding "A" as a function of the relative distance to the A prototype. The PRS hypothesis predicts that this endorsement probability should decrease more quickly with relative distance in (A, not A) tasks than in (A, B) tasks.

Of course, it makes sense to compare endorsement curves across tasks only if we somehow equate the amount of distortion within the two tasks. For this reason, we constructed the categories in such a way that the category separation was equal in each task for the two levels of distortion. To the greatest extent possible, we also tried to equate category separation across tasks. However, (A, not A) and (A, B) tasks are so fundamentally different that equating separation exactly might be impossible. For example, in an (A, not A) task the Category A exemplars are surrounded on all sides of stimulus space by "not A" category members, whereas in an (A, B) task "not A" (i.e., B) exemplars border Category A on only one side (e.g., compare Figures 2 and 3 below). We equated category separation across levels of distortion by increasing the distance between the A and B prototypes in the high-distortion condition relative to the low-distortion condition. Details are given in the next section.

EXPERIMENT 1

Method

Participants and Design

Forty-four participants from the University of California, Santa Barbara, received course credit for their participation. We used a 2 × 2 factorial design, with two different tasks [(A, not A) vs. (A, B)] crossed with two different levels of distortion (low and high). In the (A, not A) task, 9 participants participated in the low condition and 14 in the high condition. In the (A, B) task, 10 participants participated in the low condition and 11 in the high condition. Each participant participated in only one experimental condition and all participants reported 20/20 vision, or vision corrected to 20/20. Each participant completed one session that lasted approximately 25 min.

Stimuli and Stimulus Generation

A prototype distortion task was used in all conditions. Each stimulus pattern was composed of nine white circular dots displayed against a black background. Each dot could vary across trials in its horizontal (x) and vertical (y) screen positions. An entire pattern subtended a visual angle of approximately 11º, which is roughly the size of the parafovea. Because there are 9 dots, each stimulus is

Figure 1. The Category A prototype used in Experiments 1 and 2, along with typical low- and high-distortion Category A exemplars.

Table 1
Variance (in Pixels) Associated With Each Category in Experiment 1

	(A, B)			(A, not A)	
	Low	High		Low	High
A	4,715	16,570	A	4,715	16,570
B	4,715	16,570	Not A	n/a	n/a

described by 18 numerical values (i.e., 9 horizontal positions and 9 vertical positions). Examples of random dot patterns used in the task are given in Figure 1.

We begin by describing our method for generating stimuli from the (A, B) conditions.

Step 1: Construct the Category A prototype. The Category A prototype was created by randomly sampling nine (x, y) coordinates over the whole screen space (832 × 624 pixels), subject to the constraint that the mean spatial position of all dots fell exactly in the center of the screen. This eliminated the possibility that participants could use the overall spatial location of the pattern as a cue for responding.

Step 2: Create the Category A exemplars. The Category A exemplars were created by perturbing the Category A prototype by an amount that was determined by sampling from an 18-dimensional multivariate normal distribution with mean **0** and variance–covariance matrix equal to σ_L^2 I for the low-distortion condition and σ_H^2 I for the high-distortion condition (where **0** is a vector of zeros and I is the identity matrix). This algorithm is equivalent to perturbing each of the 18 horizontal and vertical dot positions by sampling from a normal distribution with mean 0 and variance either σ_L^2 or σ_H^2. The values of σ_L^2 and σ_H^2 are listed in Table 1. In all cases, if a sampled distortion produced a pattern that included dots that would be displaced off the screen, then this distortion was discarded and a new sample was selected.

Step 3: Trim the categories. All exemplars more than two standard deviations from the prototype (in 18-dimensional space) were removed from the stimulus set. This trimming served two purposes. First, it prevented any overlap in the categories, and second, it allowed us to precisely control the separation between contrasting categories.

Step 4: Construct the Category B prototype. The Category B prototype was generated by randomly sampling nine (x, y) coordinates over the whole screen space, subject to the following constraints: (1) the mean spatial position of all dots fell exactly in the center of the screen, and (2) the (Euclidean) distance (in 18-dimensional space) between the A and B prototypes equaled $4\sigma_L + \Delta$ in the low-distortion condition, or $4\sigma_H + \Delta$ in the high-distortion condition, where we used a numerical value of $\Delta = 150$ in all conditions. Note that the constant Δ, which does not depend on level of distortion, is the smallest possible distance between the nearest Category A and Category B exemplars.

Step 5: Create the Category B exemplars. The Category B exemplars were created by distorting the Category B prototype using Steps 2 and 3.

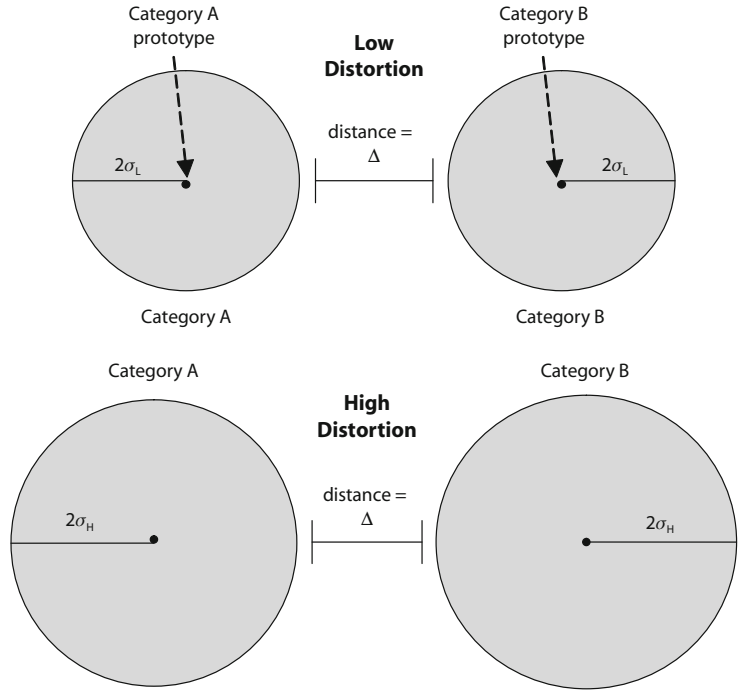

Figure 2. An illustration of the methods used to construct the Category A and B distributions used in the (A, B) tasks reported in this article (represented in 2-D).

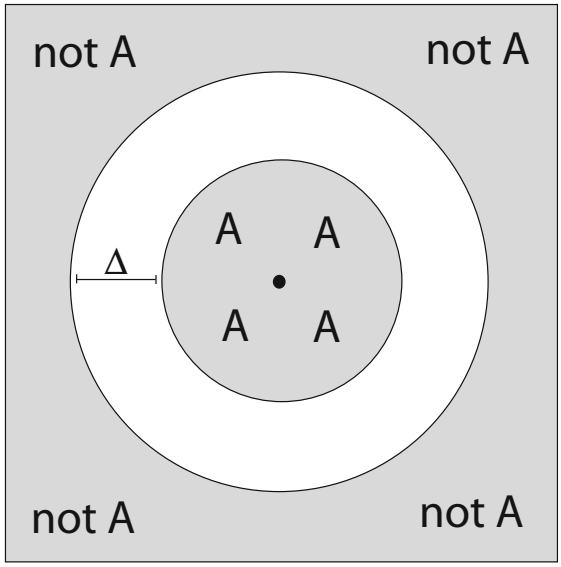

Figure 3. An illustration of the methods used to construct the Category A and "not A" distributions used in the (A, not A) tasks reported in this article (represented in 2-D).

A schematic illustrating the results of applying these steps is shown in Figure 2 (i.e., in 2 dimensions rather than in 18). Note that in both the low- and high-distortion conditions, the smallest possible distance between the nearest A and B exemplars is the same (i.e., Δ). We next created the low- and high-distortion (A, not A) conditions to preserve this minimum-distance property.

Step 6: Create the Category A exemplars for the (A, not A) conditions. We used exactly the same A categories as in the (A, B) conditions.

Step 7: Create the "not A" exemplars. The "not A" stimuli were generated by sampling exemplars uniformly over all of 18-dimensional screen space, subject to the constraint that the distance to the A prototype was greater than $2\sigma_L + \Delta$ in the low-distortion condition and greater than $2\sigma_H + \Delta$ in the high-distortion condition.

Figure 3 illustrates the (A, not A) categories. The categories constructed with this algorithm were used in both experiments reported in this article.

Procedure

The participants were run in separate cubicles on separate iMac computers in a dimly lit room. The MATLAB programming language was used to generate the visual stimuli on the screen, which was placed 35 cm from the participants. For the (A, B) conditions, participants were told that each stimulus belonged to either Group A or Group B. For the (A, not A) conditions, participants were told that each stimulus either belonged to Group A or not . Participants were instructed to indicate their response by pressing the appropriate labeled key on the keyboard. The "A" and "B" (and "A" and "not A") group labels covered the "D" and "K" keyboard keys.

Each participant took part in only one experimental condition. In all of the conditions, participants depressed the two response keys with their index fingers, and trial-by-trial feedback was provided. A brief (1-sec) high-pitched tone (500 Hz) was presented if the response was correct, and a low-pitched tone (200 Hz) was presented if the response was incorrect. Each participant completed 300 trials (10 blocks of 30). Numerical feedback was provided at the end of each block of 30 trials indicating the percentage of correct responses during that block. Participants were given 5 sec to respond. If a response was not given in the allotted time, the participant was prompted to speed up his or her response, and the trial was discarded. There was a 1-sec delay from the end of the feedback tone to the presentation of the next stimulus. Participants were naive as to the category structures; they were told that at the beginning of the experiment their responses would be guesses, but that—by using the trial-by-trial feedback provided to them—they could potentially reach 100% accuracy.

Results

We begin with standard accuracy-based analyses. Following this, we fit some popular cognitive-based categorization models to the data.

Accuracy-Based Analyses

Figure 4 shows block-by-block accuracy averaged across participants for each of the four conditions, and Table 2 shows mean accuracy across all blocks in each of the four conditions. Even though the categories in the four conditions were constructed in such a way that the nearest exemplars in the contrasting categories were approximately equidistant, Figure 4 shows that accuracy was lower in the (A, not A) conditions than in the (A, B) conditions. These differences were significant both by a sign test [low distortion, 10/10, $p < .001$; high distortion, 10/10, $p < .001$] and by a 2×2 ANOVA [$F(1,40) = 53.47$, $p < .001$].

Figure 4 and Table 2 also show that the effect of increasing distortion level was different in the (A, B) and (A, not A) tasks. In the (A, B) conditions, increasing dis-

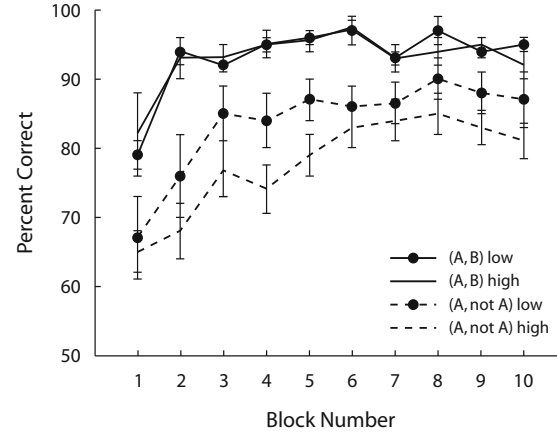

Figure 4. Block-by-block accuracy for each of the four conditions of Experiment 1.

Table 2
Mean Accuracy in Experiment 1 Across All Participants for Each of the Four Conditions

	Low		High	
	M	*SD*	*M*	*SD*
(A, B)	93.7	0.98	93.4	0.95
(A, not A)	84.0	2.52	78.1	2.03

Note—Accuracy is given by percentage of correct responses across all 10 experimental blocks.

tortion had no effect on accuracy [t test, $t(19) = 0.15$, $p > .25$; sign test, 5/10, $p > .25$]. In the (A, not A) conditions, accuracy was lower when distortion was high [t test, $t(21) = 1.82$, $p < .1$; sign test, 10/10, $p < .001$].

As described above, when testing the PRS hypothesis, the dependent measure of primary interest is the proportion of correct "A" responses as a function of the relative distance between the stimulus and the A prototype. The PRS hypothesis predicts that these endorsement curves, which are shown in Figure 5, should be steeper for the (A, not A) task. The Figure 5 curves were constructed in the following way. In the (A, not A) task, for each Category A stimulus, the distance was computed to the Category A prototype in 18-dimensional stimulus space, and this value was used to assign the stimulus to one of 6 (low-distortion) or 12 (high-distortion) distance bins. Next, within each bin, the proportion of correct "A" responses was computed. In the (A, B) task, for each Category A stimulus, the relative distance[1] was computed to the Category A prototype. These distances were grouped into either three (low-distortion) or six (high-distortion) distance bins and then the proportion of correct responses was computed for all stimuli within each bin. Note that neither the "not A" data from the (A, not A) task nor the "B" data from the (A, B) task were used in this analysis. In the case of the (A, not A) task, this is because the PRS hypothesis does not make strong predictions about how "not A" stimuli will be classified. It does make strong predictions about "B" trials in the (A, B) task, but correct "B" responses from the (A, B) task do not serve as a proper comparison with correct "A" responses from the (A, not A) task (since the stimuli were different). The "A" stimuli were identical in the (A, not A) and (A, B) tasks, so it is most appropriate to compare correct "A" responses across tasks.

First, note that the (A, B) endorsement curves contain fewer points than the (A, not A) curves. This is because of the inherent difficulty difference between the tasks. In the

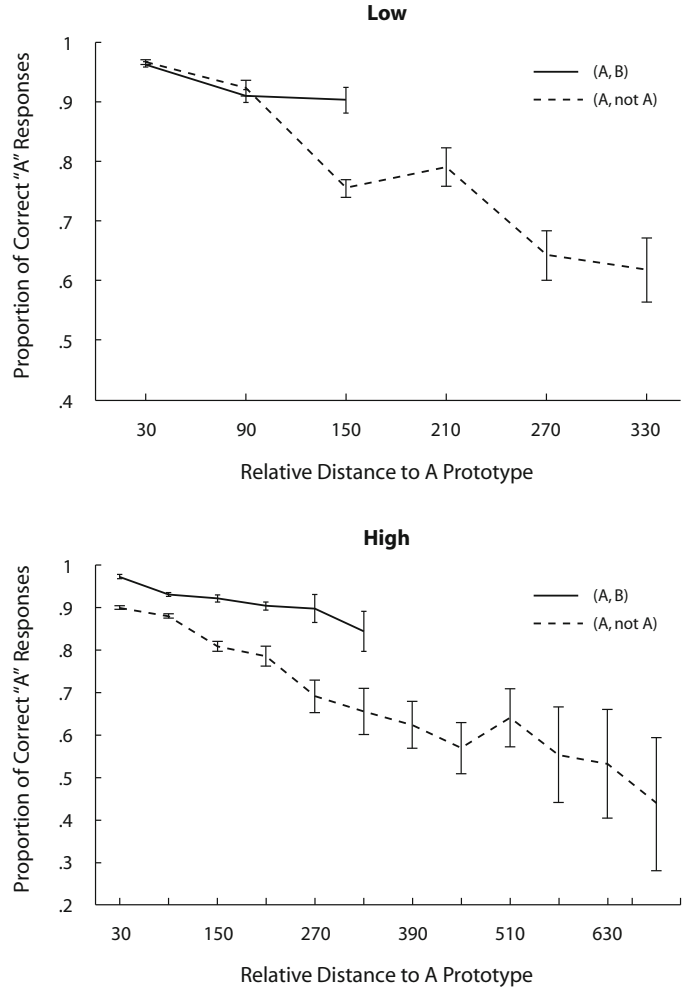

Figure 5. The proportion of correct "A" responses in Experiment 1 for both the (A, B) and (A, not A) low-distortion (top panel) and high-distortion (bottom panel) conditions, as a function of relative distance to the A prototype.

(A, not A) task, there are many more Category A exemplars near exemplars in the contrasting category than in the (A, B) task, and as a result there are more stimuli of high difficulty. Second, note that a visual inspection of Figure 5 appears to support the PRS prediction that the endorsement curves are steeper for the (A, not A) task than for the (A, B) task.

To test this prediction statistically, the endorsement data from the first three relative distances of the low-distortion conditions were subjected to a 2 × 3 ANOVA {2 tasks [(A, not A) vs. (A, B)] × 3 relative distances from the A prototype} and the data from the first six relative distances of the high-distortion conditions were subjected to a 2 × 6 ANOVA {2 tasks [(A, not A) vs. (A, B)] × 6 relative distances from the A prototype}. For the high level of distortion, there was a significant interaction between task type [(A, not A) vs. (A, B)] and relative distance [low distortion: $F(5,138) = 3.30$, $p < .01$], which supports the conclusion that the proportion of "A" responses fell off more quickly in the (A, not A) condition than in the (A, B) condition. In the low-distortion condition, the interaction, readily apparent in Figure 5, was not significant [$F(2,51) = 0.63$, $p > .05$]. In addition, there was a main effect of task type [(A, not A) vs. (A, B)] in the high- [$F(1,138) = 37.95$, $p < .001$] but not in the low- [$F(1,51) = 1.4$, $p > .05$] distortion conditions, which confirms our conclusion that participants found the (A, not A) conditions more difficult than they found the (A, B) conditions. Finally, there was also a (less interesting) main effect of distance from prototype [low distortion, $F(2,51) = 10.97$, $p < .001$; high distortion, $F(5,138) = 6.41$, $p < .001$].

A visual inspection of Figure 5 seems to indicate that the steepness of the (A, not A) endorsement curve relative to the (A, B) curve is greater in the low-distortion condition than in the high-distortion condition. Thus, the failure of this interaction to reach significance in the low-distortion condition is somewhat surprising. There are several possibilities. First, in the low-distortion condition, the ANOVA included only three levels of distance, whereas in the high-distortion condition there were six levels of distance. As a result, the low-distortion ANOVA had fewer degrees of freedom and less power than the high-distortion ANOVA. Second, accuracy was very high at the two lower distances in both tasks of the low-distortion condition, and this ceiling effect may have obscured an interaction at the lower distances. There is some statistical support for this hypothesis. First, note that at the smallest distance, the (A, not A) and (A, B) proportions are virtually identical in the low-distortion condition. In contrast, at the third distance level—that is, the largest (A, B) distance—the (A, not A) proportion is significantly lower than is the (A, B) proportion [$t(17) = 9.114$, $p < .001$]. In summary, there appears to be a ceiling effect in the low-distortion condition. In general, ceiling effects reduce interactions, and this seems to be the case here.

Model-Based Analyses

The PRS hypothesis predicts that participants will use different strategies in (A, not A) and (A, B) prototype distortion tasks. Before concluding that our results support this hypothesis, it is important to determine whether models that assume that the same cognitive strategy is used in both tasks are compatible with our results. The two most popular cognitive accounts of prototype distortion performance are from prototype theory and exemplar theory.

Prototype theory assumes that when an unfamiliar stimulus is encountered, it is assigned to the category with the most similar prototype (e.g., Homa et al., 1981; Posner & Keele, 1968, 1970; Reed, 1972; Rosch, 1973, 1975; Smith & Minda, 1998). In prototype distortion tasks, the prototype model is equivalent to the ideal observer.[2] Exemplar theory assumes that, when an unfamiliar stimulus is encountered, its similarity is computed to the memory representation of every previously seen exemplar from each potentially relevant category. The probability that the stimulus is assigned to a particular category depends on the relative magnitude of the sum of all similarities associated with that category (Brooks, 1978; Estes, 1986, 1994; Hintzman, 1986; Lamberts, 2000; Medin & Schaffer, 1978; Nosofsky, 1986).

Let D_{xA} denote the dissimilarity or psychological distance from stimulus x to the Category A prototype, and define D_{xB} analogously. Then, according to the prototype model, the probability of responding "A" in an (A, B) task equals

$$P_{(A,B)}(A|x) = P\left(D_{xB} - D_{xA} > \varepsilon\right) = P\left(Z < \frac{D_{xB} - D_{xA}}{\sigma}\right), \quad (1)$$

where ε is normally distributed with mean $\mathbf{0}$ and variance σ^2. The (A, not A) task is similar, except D_{xA} is compared with a threshold T rather than with D_{xB}. Specifically,

$$P_{(A,\text{not A})}(A|x) = P\left(T - D_{xA} > \varepsilon\right) = P\left(Z < \frac{T - D_{xA}}{\sigma}\right). \quad (2)$$

The exemplar model bases category decisions on the sum of the similarities of the stimulus to all exemplars of the relevant contrasting categories. Let S_{xi} denote the similarity of stimulus x to previously seen exemplar i. Then, according to the exemplar model, the probability of responding "A" in an (A, B) task (Ashby & Maddox, 1993; Medin & Schaffer, 1978; Nosofsky, 1986) equals

$$P_{(A,B)}(A|x) = \frac{\left(\sum_{i \in A} S_{xi}\right)^{\gamma}}{\left(\sum_{i \in A} S_{xi}\right)^{\gamma} + \left(\sum_{j \in B} S_{xj}\right)^{\gamma}}. \quad (3)$$

The parameter γ is a measure of response variability. For example, it is inversely related to the value of σ in Equations 1 and 2 (Ashby & Maddox, 1993). In (A, not A) tasks, the summed similarities to Category A exemplars are compared with a threshold T (e.g., Nosofsky & Zaki, 1998),

$$P_{(A,\text{not A})}(A|x) = \frac{\left(\sum_{i \in A} S_{xi}\right)^{\gamma}}{\left(\sum_{i \in A} S_{xi}\right)^{\gamma} + T}. \quad (4)$$

Following Posner, Goldsmith, and Welton (1967) and Smith and Minda (2001), we assumed

$$D_{xi} = \log(1 + Distance_{xi}), \quad (5)$$

where $Distance_{xi}$ is the Euclidean distance in 18-dimensional stimulus space between stimulus x and exemplar i. Following Shepard (1987), we assumed that similarity and psychological distance are related via

$$S_{xi} = \exp(-D_{xi}). \quad (6)$$

The PRS hypothesis predicts that when the PRS by itself is used to select a response in the (A, not A) task, the decision about whether to respond "Yes" will depend on how familiar the stimulus feels. At the computational level, this is similar to the decision strategy assumed by prototype theory. Because the prototype is the most likely stimulus in either category,[3] it should elicit the greatest feeling of familiarity. As distance from the prototype increases, likelihood decreases and, therefore, so should familiarity.

In contrast, the decision strategy assumed by the exemplar model is less compatible with the decision strategy assumed by the PRS. For example, the exemplar model predicts that a stimulus from Category A that is highly similar to a previously seen atypical Category A exemplar is likely to elicit a "Yes" response. Such a stimulus should elicit some PRS activation, and therefore seem vaguely familiar. However, it is important to note that a feeling of vague familiarity could also occur on "not A" trials. Even though within-category similarity is low in the "not A" category, some pairs of "not A" category members will be highly similar, by chance. Thus, a feeling that a stimulus is vaguely familiar is not enough to signal a participant to respond "Yes." As a result, a pure PRS strategy predicts that participants should respond "No" to all atypical Category A exemplars, whereas the exemplar model predicts that some of these stimuli should elicit a "Yes" response. The predictions of a pure PRS strategy and the exemplar model therefore disagree.[4]

The PRS hypothesis makes a number of specific predictions about how well the prototype and exemplar models should fit the Experiment 1 endorsement curves. First, as noted above, because the PRS hypothesis assumes that multiple systems are used, it predicts that both models should make systematic mispredictions—that is, since they both assume a single system. Second, it predicts that the empirical (A, not A) endorsement curves should be steeper than predicted by the exemplar model—that is, since it predicts that the PRS will mediate many (A, not A) responses. Third, it predicts that the empirical (A, B) endorsement curves should be shallower than predicted by the prototype model—that is, since the PRS by itself is insufficient in the (A, B) task.

In the present application, the prototype and exemplar models both have three similar parameters. In both models there is one threshold parameter T for the low-distortion (A, not A) condition, and a separate T for the high-distortion (A, not A) condition. The remaining free parameter for the prototype model is σ^2, which is a measure of the magnitude of perceptual noise. Since the stimuli were similar in all four conditions, we assumed that σ^2 did not vary with task or level of distortion. For the exemplar model, a similar role is played by the γ parameter, which is inversely related to perceptual noise (Ashby & Maddox, 1993).

As is standard for these models, we assumed that the prototypes used by the prototype model were the true category prototypes and that the similarities in the exemplar model were computed to all category members. We also more crudely fit versions of the models that were sensitive to the stimulus presentation order. In this version of the prototype model, the Category A prototype is estimated by computing the mean of all previously seen Category A exemplars. As a result, the Category A prototype is updated after every Category A trial. In the dynamic version of the exemplar model the summed similarities in Equations 3 and 4 are computed only for those category exemplars already in memory. Thus, after every trial, another exemplar is stored in memory and on the following trial one more term is added to one of the sums. These more complex versions of the prototype and exemplar models made the same qualitative predictions as the simpler static versions, so we do not discuss them further.

Equations 1–4 were used to generate the predicted probability of responding correctly to every stimulus in each of the four experimental conditions of Experiment 1 for the prototype and exemplar models. These probabilities were then grouped according to the (Euclidean) distance from each stimulus to the A prototype in the (A, not A) conditions (in 18-dimensional stimulus space) and according to the relative distance $D_{xA} - D_{xB}$, in the (A, B) conditions. For both models, the parameters were estimated using the method of maximum likelihood. The two models each have three free parameters, so there is no reason to use a penalized measure of fit.

Figure 6 shows the best fits of the prototype model and Figure 7 shows the best fits of the exemplar model to the data from all four conditions of Experiment 1. Note first that all conditions contain extra data not shown in Figure 5. In the (A, not A) conditions, these are the proportion of correct "not A" responses, and in the (A, B) conditions, these are the proportion of correct "B" responses. For the reasons mentioned above, these were excluded from Figure 5. Second, note that both models provide slightly better fits to the (A, not A) data than they do to the (A, B) data. This is at least partly due to the fact that both models had two free threshold parameters (i.e., T in Equations 2 and 4) that could be manipulated to improve the fits to the (A, not A) data, whereas there were no analogous parameters in the (A, B) conditions. Thus, both models had greater mathematical flexibility in the (A, not A) than in the (A, B) conditions. As a result, the better fits to the (A, not A) data do not necessarily mean that the models have greater psychological validity in (A, not A) conditions.

Table 3 lists the overall goodness-of-fit values ($-2 \ln L$) for each model, as well as the fit values for each of the four conditions. The values of the best-fitting parameters are listed in the Figures 6 and 7 figure captions. Note from Table 3 that the prototype model provided the better fit in both (A, not A) conditions. In the (A, B) conditions, the fits of the two models were almost equal, although in both cases the exemplar model had a slight advantage. Overall, however, the prototype model provided a slightly better fit.

These fits should be interpreted with some caution, because they depend on our assumptions about how psycho-

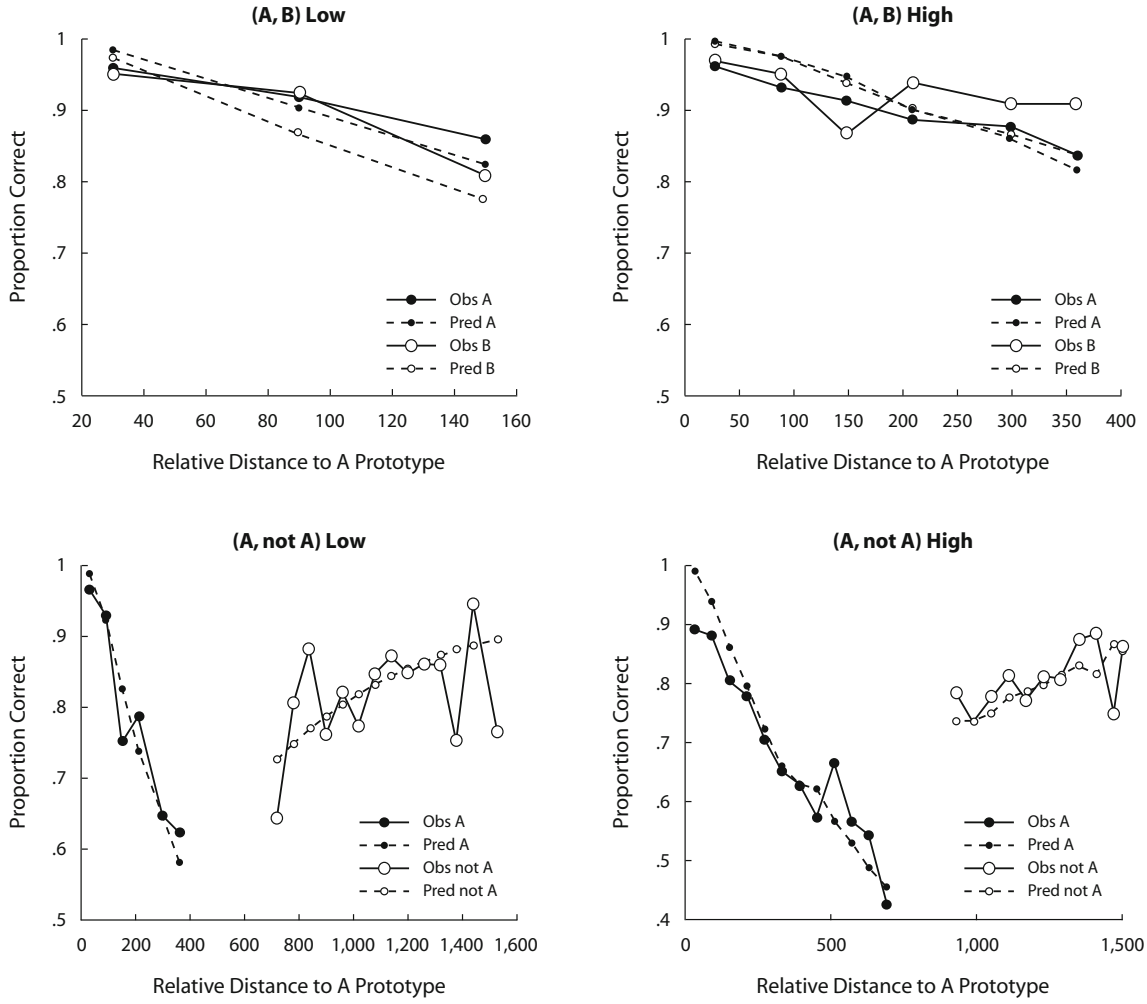

Figure 6. Prototype model fits to the Experiment 1 data shown in Figure 5. The best-fitting value of the standard deviation of perceptual noise (σ) was 1.0613, and the best-fitting values for the thresholds (T) were 6.006 for the low-distortion condition and 6.429 for the high-distortion condition.

logical distance and similarity are related to physical distance in the 18-dimensional stimulus space (see Equations 5 and 6). The assumptions we made are the most common choices (i.e., we know of no alternatives to Equation 5), but it is possible that some other method of measuring similarity might have qualitatively changed these fits. Another problem with applying the models is that they both require some solution to the dot correspondence problem. As mentioned above, each stimulus is described by 18 numerical values. Values 1 and 2 are the horizontal and vertical screen coordinates of the first dot, values 3 and 4 are the horizontal and vertical screen coordinates of the second dot, and so forth. The distance between two patterns in stimulus space is computed by comparing the 18 numbers describing the first pattern with the corresponding 18 numbers that describe the second pattern. But this process assumes that we know how to number the dots from 1 to 9 in each pattern (so that dot 1 in the first pattern is compared with dot 1 in the second pattern). If the two patterns are identical, or are both low distortions of the Category A prototype, the correspondence is obvious. But suppose that we wish to compute the distance between two "not A" patterns: In this case, the dots will have no obvious correspondence, and there will be 362,880 possible solutions to the correspondence problem (i.e., 9!). It would clearly be impossible to collect empirical evidence to decide which solution is best,

Table 3
Goodness-of-Fit Values ($-2 \ln L$) for the Prototype and Exemplar Models in Experiment 1

	Prototype	Exemplar
(A, B) low	2.93	2.91
(A, not A) low	5.63	5.72
(A, B) high	24.18	24.17
(A, not A) high	15.51	16.58
Overall	48.25	49.38

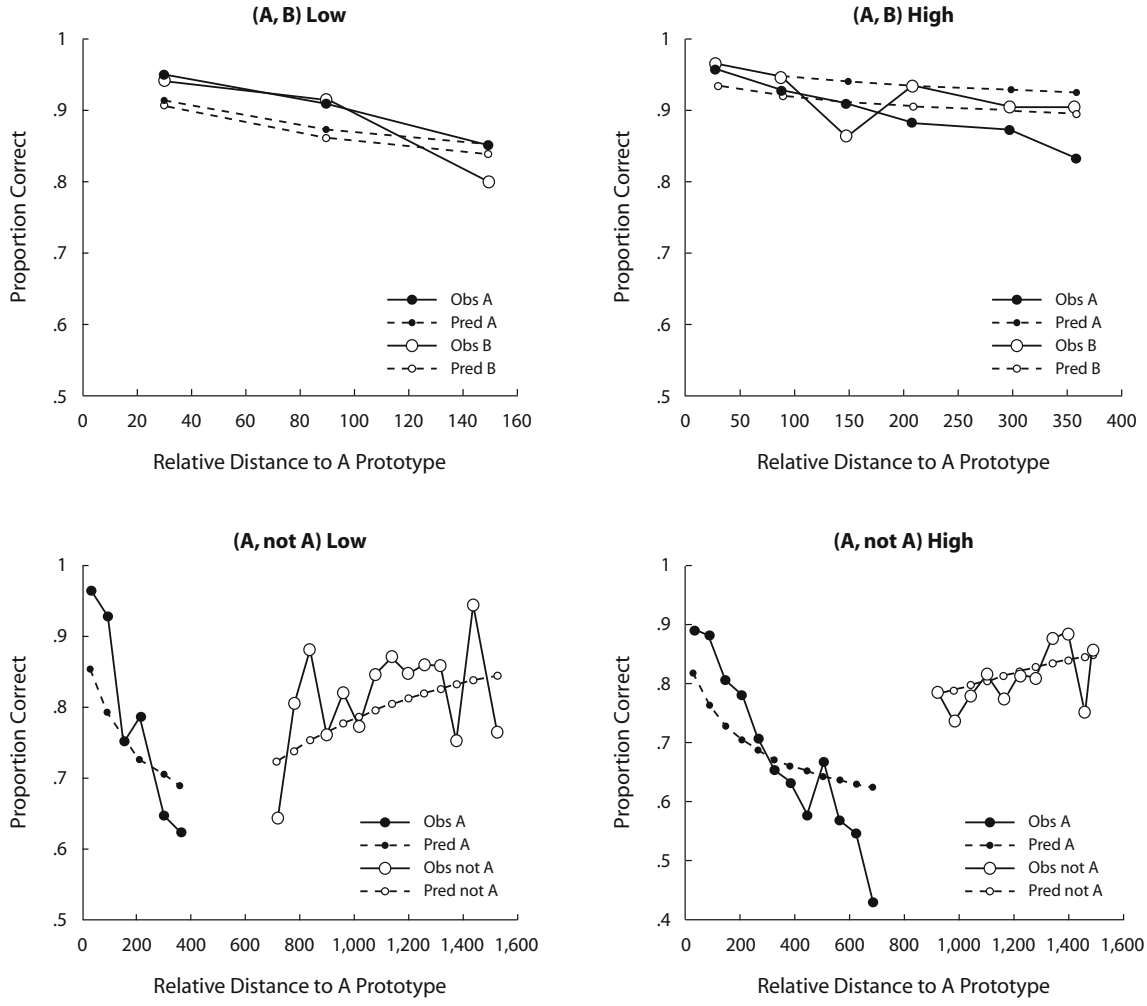

Figure 7. Exemplar model fits to the Experiment 1 data shown in Figure 5. The best-fitting value of the gamma parameter (γ) was 1.1088, and the best-fitting values for the thresholds (T) were 0.8507 for the low-distortion condition and 0.8585 for the high-distortion condition.

so any choice is arbitrary.[5] Fortunately, there are mathematical reasons to believe that this choice will have little effect on the model fits (see our discussion of stimulus dimensionality in the General Discussion). We chose the null solution,[6] but it is impossible to know whether one of the other 362,880 possible solutions might have changed the outcome of this fitting process.

As described above, the PRS hypothesis predicts that the prototype and exemplar models should both make certain systematic mispredictions when fit to the empirical endorsement curves. These predictions are most easily tested by examining the residuals associated with each model—that is, by examining the difference between the predicted probability and the observed proportion of responses at each relative distance. These residuals are shown in Figure 8 for each of the four conditions, along with the best-fitting regression line and the correlation between residual and distance. In the (A, not A) task, the residuals are plotted only for Category A exemplars (i.e., again because the PRS hypothesis makes no predictions about "not A" trials).

Increasing regression lines (and positive correlations) mean that the empirical endorsement curves are steeper than predicted in a model, and decreasing regression lines (and negative correlations) mean that the empirical endorsement curves are shallower than predicted in a model. Note that all three predictions of the PRS hypothesis are supported in Figure 8. First, both models systematically mispredict the empirical endorsement curves—the prototype model predicts curves that are too steep and the exemplar model predicts curves that are too shallow. Second, the empirical (A, not A) endorsement curves are steeper than predicted by the exemplar model. Third, the empirical (A, B) endorsement curves are shallower than predicted by the prototype model.

Note that our findings in the (A, not A) conditions replicate the results of Smith and Minda (2001), who showed that

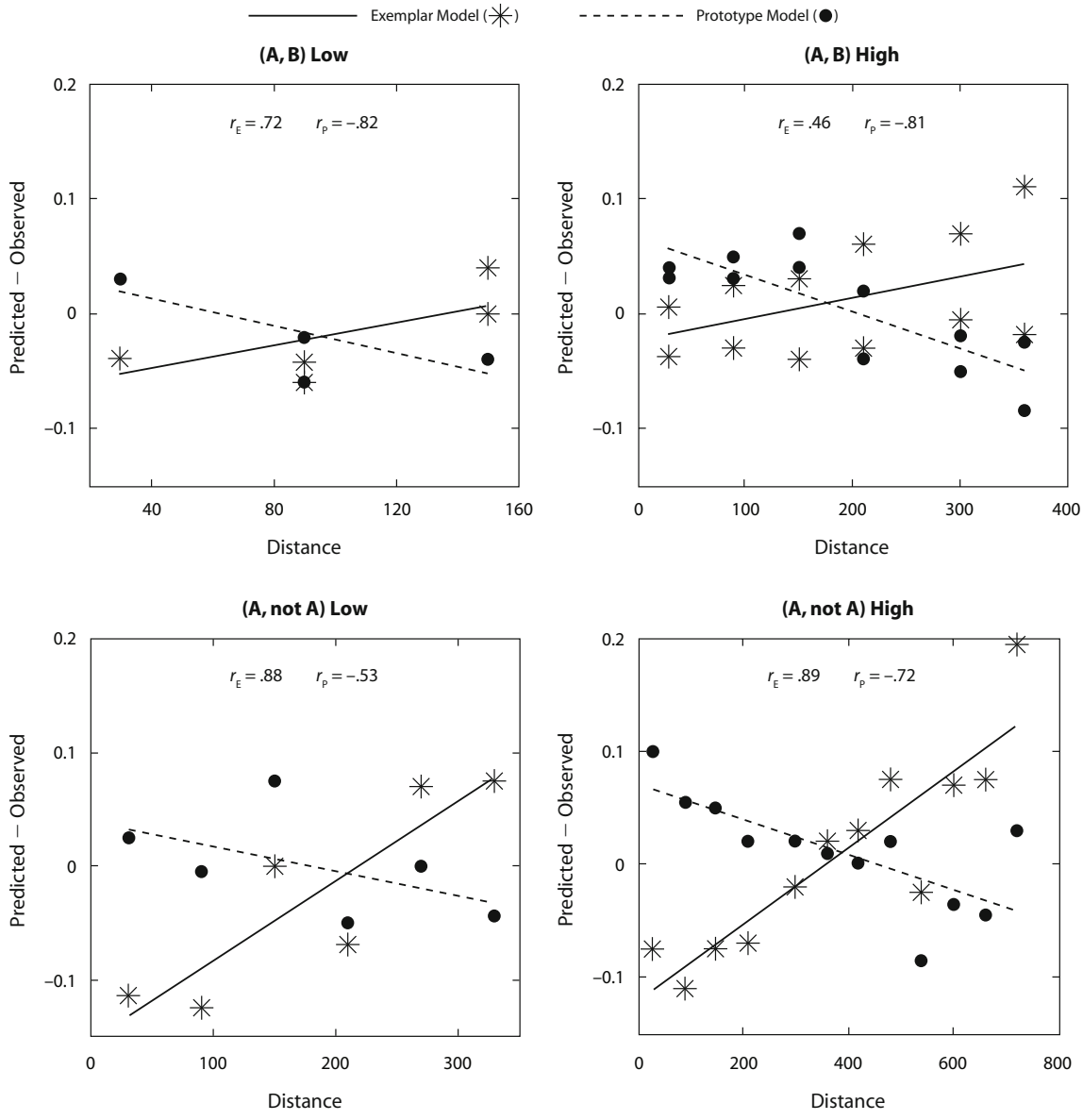

Figure 8. Residuals (i.e., predicted probability − observed proportion) for the exemplar and prototype models in each condition of Experiment 1. The (A, not A) figures only include A trials. Also shown are the best-fitting lines and the correlation coefficients—that is, between the residual and relative distance to the prototype (r_E = correlation for exemplar model; r_P = correlation for prototype model).

the exemplar model predicted endorsement curves that were shallower than the curves they observed in their (A, not A) conditions. Zaki and Nosofsky (2004) reported that when participants were trained on only high-level distortions, the (A, not A) endorsement curves became less steep. Note that this result is also predicted by the PRS hypothesis. When training only entails high-level distortions, activation of the PRS by Category A exemplars will be low, and as a result, the endorsement curves should be less steep.

Figure 8 also argues against the hypothesis that our main conclusions in Experiment 1 were driven by ceiling effects in the (A, B) conditions. It is true that (A, B) accuracy was high at both levels of distortion, but note that the exemplar and prototype models both systematically mispredicted the (A, B) results, and that these mispredictions were qualitatively the same as in the (A, not A) conditions. Both models can perfectly fit data in which accuracy is perfect. Yet the exemplar model predicted endorsement curves that were too shallow in the (A, B) conditions, and the prototype model predicted curves that were too steep; (A, B) accuracy was, therefore, not too high to differentiate between these models.

Discussion

The results clearly show that the proportion of correct "A" responses decreased with the relative distance between the presented stimulus and the Category A prototype more quickly in the (A, not A) task than in the (A, B) task. The results from these two tasks also differed qualitatively in other ways. First, even though the distance between the nearest exemplars in contrasting categories was approximately equal in the two tasks, participants performed more poorly in the (A, not A) tasks than in the (A, B) tasks. Second, overall accuracy was invariant across distortion levels in the (A, B) task (see Figure 4), suggesting that participants dealt with distortion optimally. In contrast, (A, not A) performance deteriorated when distortion was increased. Note that the PRS hypothesis predicts that increasing distortion levels should be more detrimental in the (A, not A) task because when distortion is increased, fewer Category A members will be similar to the A prototype; and, therefore, fewer trials will induce strong PRS activation.

The two leading cognitive accounts of prototype distortion tasks—prototype and exemplar theories—both provided good quantitative fits to the Experiment 1 data. Even so, both models made systematic mispredictions. In particular, the prototype model predicted endorsement curves that were too steep and the exemplar model predicted endorsement curves that were too shallow.

The PRS hypothesis correctly predicted that the (A, not A) endorsement curves were steeper than the (A, B) curves, that the exemplar model (A, not A) endorsement curves would be too shallow, and the prototype model (A, B) curves would be too steep. At this stage of its development, the PRS hypothesis does not make specific quantitative predictions. Therefore, it should not be viewed as an alternative to prototype or exemplar models. In this sense, it is somewhat like Nosofsky's (1986; Nosofsky & Johansen, 2000) attention optimization hypothesis, which also does not make precise quantitative predictions, but nevertheless does make qualitative predictions about what should happen when quantitative models are fit to categorization data.

EXPERIMENT 2

The results of Experiment 1 provide the first known behavioral evidence supporting a role for the PRS in category learning. They verified a prediction of the PRS hypothesis that the probability of correctly responding "A" should decrease with distance to the A prototype more quickly in (A, not A) tasks than in (A, B) tasks. Despite the simplicity of this hypothesis, however, it also makes other strong predictions. For example, the PRS does not depend on feedback for learning, simple repetition is sufficient (e.g., Schacter, 1990; Wiggs & Martin, 1998). Therefore, if the PRS plays a key role in (A, not A) tasks, then performance in these tasks should not depend critically on trial-by-trial feedback. In contrast, we have argued that learning in (A, B) tasks must be mediated primarily by memory systems other than the PRS. Our results do not identify these other systems, but the systems most commonly thought to contribute to category learning (e.g., an explicit reasoning system, a procedural-learning system) either benefit from or require trial-by-trial feedback (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby, Queller, & Berretty, 1999). Thus, we expect that performance in (A, B) tasks should be affected by the loss of trial-by-trial feedback much more than performance in (A, not A) tasks.

Experiment 2 tested this prediction—namely, that eliminating the signal indicating whether each categorization response is "correct" or "incorrect" should be more deleterious to (A, B) performance than to (A, not A) performance. The category structures and methods were identical to those used in Experiment 1, with the single difference that, in Experiment 2, there was no trial-by-trial feedback. The use of identical methods in the two experiments allows us to use the Experiment 1 results as a full feedback control for Experiment 2.

Method

Participants and Design

Forty-one participants from the University of California, Santa Barbara, received course credit for their participation. We used a 2×2 factorial design, with two different tasks [(A, not A) vs. (A, B)] crossed with two levels of distortion (low vs. high). In the (A, not A) task, 8 participants participated in the low-distortion condition and 10 in the high-distortion condition. In the (A, B) task, 8 participants participated in the low-distortion condition and 15 in the high-distortion condition. Each participant participated in only one condition; all participants reported 20/20 vision, or vision corrected to 20/20. Each participant completed one session that lasted approximately 25 min.

Procedure

The methods used in Experiment 2 were identical to those used in Experiment 1, except that trial-by-trial feedback was removed from the task. Block feedback, however, was provided—that is, after every 30 trials, participants were informed of their percentage of correct responses on the preceding 30 trials. The instructions given to participants in Experiment 2 were similar to the instructions given to participants in Experiment 1, except they were told they would be provided no information as to whether each of their responses was correct or incorrect. In addition, participants were told that they should not change their categorization strategy after they felt confident that they had learned the categories. Participants in the (A, B) conditions were also told that it did not matter which response key they used for Category A, and which key they used for Category B. However, they were encouraged to use the same category/response key mapping throughout each block of the experiment.

Results

The absence of feedback meant the assignment of responses to buttons was arbitrary. Therefore, for each block, we computed the percentage of correct responses with each assignment and assumed that participants used the assignment for which accuracy was above chance. Overall accuracy is shown in Table 4. Note that the results are very different from Experiment 1, where performance was much better in the (A, B) conditions (see Table 2). In Experiment 2, performance was slightly better in the (A, not A) conditions (more on this below). Thus, removing feedback led to much larger drops in accuracy in the (A, B) conditions than in the (A, not A) conditions. When distortion was low, overall accuracy dropped from Experi-

Table 4
Mean Accuracy in Experiment 2 Across All Participants for Each of the Four Conditions

	Low		High	
	M	*SD*	*M*	*SD*
(A, B)	84.8	1.25	70.8	1.38
(A, not A)	88.9	0.73	69.4	1.20

Note—Accuracy is given by percentage of correct responses across all 10 experimental blocks.

ment 1 to Experiment 2 by 9% in the (A, B) condition, and it actually increased in the (A, not A) condition (i.e., by 5%). When distortion was high, removing feedback caused a 22% drop in accuracy in the (A, B) condition, compared with a 9% drop in the (A, not A) condition.

Figure 9 shows block-by-block learning curves from each of the four conditions, together with the learning curves from the corresponding conditions of Experiment 1. The top panel of Figure 9 shows data from the (A, not A) conditions and the bottom panel shows data from the (A, B) conditions. Note first that in the low-distortion condition of the (A, not A) task, accuracy consistently increases across blocks in the absence of any feedback. This is especially critical because learning that depends on the PRS should not require feedback, but it does require

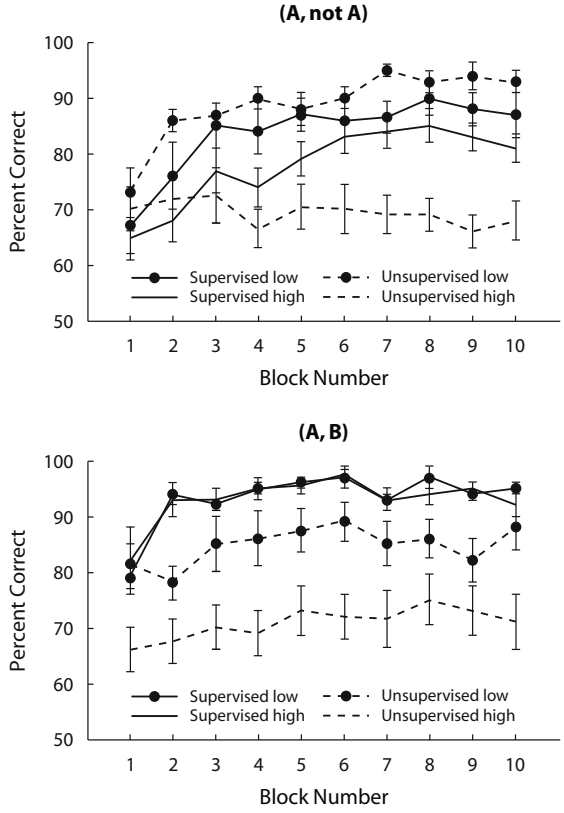

Figure 9. Block-by-block accuracy in Experiments 1 and 2. The top panel shows results for all (A, not A) conditions and the bottom panel shows results for all (A, B) conditions.

repetition. Therefore, the PRS hypothesis correctly predicts the increasing nature of the low-distortion (A, not A) learning curve in Experiment 2.

Second, note that in the low-distortion (A, not A) conditions, Experiment 2 performance is consistently better than Experiment 1 performance. Thus, removing feedback in this condition actually improved performance (this difference is significant by a sign test: 10/10, $p < .001$). In all other conditions, however, accuracy is considerably lower in the absence of feedback. An ANOVA on the (A, not A) data indicated a main effect of distortion [$F(1,39) = 29.02$, $p < .001$] and an interaction between feedback and distortion [$F(1,39) = 10.23$, $p < .01$], but no main effect of feedback [$F(1,39) = 2.57$, $p > .1$]. In contrast, for the (A, B) data, the main effects of feedback [$F(1,42) = 75.90$, $p < .001$] and distortion [$F(1,42) = 28.00$, $p < .001$] were both significant, but the interaction between feedback and distortion was not. These results strongly suggest that feedback was more critical to performance in the (A, B) conditions than in the (A, not A) conditions.

Recall that the absence of feedback meant that the assignment of responses to buttons was arbitrary. Therefore, for each block, we computed the percentage of correct responses with each assignment and assumed that participants used the assignment for which accuracy was above chance. This procedure, although necessary, slightly overestimates the true accuracy, since it guarantees that guessing can never produce accuracy below chance. So one possibility is that the higher (A, not A) accuracy without feedback in the low-distortion condition is because of this slight bias in our estimation procedure. To test this hypothesis, we reanalyzed the data, assuming that each participant used the same assignment of responses to buttons throughout the experiment. This did not change the conclusion that (A, B) performance was significantly worse without feedback, but under this assumption there no longer was a significant difference between low-distortion (A, not A) performance with and without feedback.

Another result of interest—briefly mentioned above and visible in Table 4 and by comparing the top and bottom panels of Figure 9—is that when distortion was low, unsupervised performance was generally better in the (A, not A) task than in the (A, B) task. This difference is not significant with ANOVA, but it is significant by a sign test (9/10, $p = .011$). Recall that this is the opposite of the pattern seen in Experiment 1. With feedback, the (A, not A) task was more difficult. When distortion was high, there was no significant difference between (A, not A) and (A, B) performance (i.e., either by ANOVA or a sign test).

In summary, the results of Experiment 2 strongly support the prediction of the PRS hypothesis that feedback is more important in the (A, B) task than in the (A, not A) task. The prototype and exemplar models do not make predictions in unsupervised experiments, so neither model can be fit to the Experiment 2 data.

Discussion

Experiment 2 showed that, as predicted by the PRS hypothesis, removing trial-by-trial feedback had a much greater effect on performance in the (A, B) task than in

the (A, not A) task. In the (A, not A) task, overall accuracy was only about 2% lower in the absence of feedback, whereas (A, B) accuracy dropped by an average of 16% when feedback was removed. In the (A, B) task, large accuracy drops were seen at both levels of distortion (9% in the low-distortion condition and 22% in the high-distortion condition), but in the (A, not A) task, accuracy only dropped in the high-distortion condition (by 9%). In the low-distortion (A, not A) task, accuracy actually improved by about 5% when feedback was removed. This is important, because the PRS hypothesis predicts that the effects of the PRS on performance should be greatest in the low-distortion (A, not A) condition.

GENERAL DISCUSSION

Because the A categories were identical in the (A, not A) and (A, B) conditions, participants received identical Category A training in the two tasks. In addition, category separation was approximately equal in all conditions of both experiments. Despite these similarities, (A, not A) and (A, B) performance differed qualitatively in Experiments 1 and 2: (1) when feedback was provided (A, B) performance was unaffected by distortion level, whereas (A, not A) performance deteriorated with increasing levels of distortion; (2) (A, not A) performance was significantly more sensitive to prototype similarity than was (A, B) performance; (3) (A, not A) performance was worse than (A, B) performance when trial-by-trial feedback was given, but equal or better when training was unsupervised; and (4) when distortion was low (A, not A) performance actually improved when feedback was removed, whereas (A, B) performance deteriorated. Collectively, these differences argue strongly that learning in (A, not A) and (A, B) prototype distortion tasks is mediated by functionally separate systems.

The PRS hypothesis correctly predicted most of these results a priori. In particular, it predicted that (A, not A) performance would worsen as distortion level increased. It predicted that (A, not A) performance would be more sensitive to relative distance to the A prototype than (A, B) performance would be, and it predicted that removing trial-by-trial feedback would harm (A, B) performance more than it would harm (A, not A) performance. In addition, our modeling results showed that the two most popular cognitive theories of learning in prototype distortion tasks—namely, prototype theory and exemplar theory—provided good quantitative fits to the Experiment 1 data, but at the same time they both made certain systematic mispredictions. In particular, Figure 8 shows that neither model could account for the observed steepness of the (A, not A) and (A, B) curves. The prototype model predicted curves that were too steep and the exemplar model predicted curves that were too shallow. Finally, we know of no unsupervised versions of prototype or exemplar theory, so neither model appears to make any a priori predictions about Experiment 2.

An alternative account of our results might be that in the (A, not A) conditions, participants learned only a single category (i.e., Category A), whereas in the (A, B) conditions they learned two. This hypothesis might be used, for example, to explain why learning without feedback was more successful in Experiment 2 in the (A, not A) conditions. We believe that this possibility is unlikely for two reasons, one logical and one empirical. First, there is no logical reason participants should adopt this strategy. In both the (A, not A) and (A, B) tasks, there are two categories of stimuli, each with its own separate response; and on every trial, feedback was given signaling the category membership of the stimulus on that trial. So, if participants learned only one category in the (A, not A) conditions, there is no reason not to adopt this same strategy in the (A, B) conditions. Second, this hypothesis predicts that there should have been a main effect of feedback when comparing Experiments 1 and 2, not an interaction; in other words, if participants in Experiment 2 found that, in the absence of feedback, the (A, not A) task was less difficult than the (A, B) task, because they only had one category to learn in the (A, not A) task but two in the (A, B) task, this same ordering of task difficulty should have been seen in Experiment 1 when feedback was provided on every trial. But instead, the opposite result was observed—that is, under supervised conditions, the (A, B) task was easier.

As it happens, there is another feature of the stimuli used in the present experiment that might facilitate the use of the PRS in the (A, not A) conditions—namely, that the dot patterns vary on many perceptual dimensions rather than on few. We have hypothesized that PRS activation in the prototype distortion task requires many category exemplars to be similar to the prototype. In the prototype distortion task, every exemplar is a unique distortion of the prototype. Therefore, the PRS should be more active if random distortions produce many stimuli similar to the prototype. With stimuli that vary on only one dimension, random distortions of a prototype will produce some exemplars with a lower value than the prototype on the stimulus dimension and some with a higher value. As a result, a few distortions will be close to the prototype (and therefore similar) and many will be further away (and therefore more dissimilar). In fact, in one dimension, only two exemplars can be the nearest neighbors of the prototype. All other exemplars must be more dissimilar to the prototype than these two can be. In two dimensions, however, five exemplars can be the nearest neighbor of the prototype, because now the exemplars can cluster around the prototype at all compass points instead of simply falling to the left or right. As stimulus dimensionality increases, this trend accelerates. For example, with 8-dimensional stimuli, 240 different exemplars can all be nearest neighbors of the prototype, and with stimuli that vary on 24 dimensions, the number of possible nearest neighbors of the prototype increases to 196,560 (Odlyzko & Sloane, 1979). Thus, random distortions of the prototype are likely to produce more exemplars highly similar to the prototype when the stimuli vary on many perceptual dimensions. With nine dots, the random dot patterns vary on 18 stimulus dimensions. As a result, many distortions of the prototype will lie very near the prototype in stimulus space and will likely activate the PRS.[7] Thus, the PRS hypothesis predicts that the difference between (A, not A) and (A, B) learning observed in the present study should be less pronounced with stimuli that vary on fewer dimensions.

The neural basis of the PRS is still unclear. The repetition priming thought to be mediated by the PRS is widely associated with the phenomenon of repetition suppression, in which repeated presentations of a stimulus elicit a smaller and smaller neural response (e.g., Raichle et al., 1994; Schacter & Buckner, 1998; Wiggs & Martin, 1998). As a result, there have been specific proposals that repetition suppression is the neural signature of PRS activation (e.g., Schacter & Buckner, 1998; Wiggs & Martin, 1998).

Linking PRS activation to repetition suppression should facilitate the development of a neural theory of the PRS, but several important questions remain unanswered. First, if the PRS is a purely perceptual memory system, we might expect to see its effects limited to sensory areas of cortex (including sensory association areas). It is true that repetition suppression is often seen in visual cortex, but it has also been reported in other nonsensory brain areas, including prefrontal cortex (e.g., Demb et al., 1995; Raichle et al., 1994; Wagner, Desmond, Demb, Glover, & Gabrieli, 1997). Second, the neural mechanisms that mediate repetition suppression are also unclear. For example, it is unclear whether repetition suppression is due to a sharpening of tuning curves (Wiggs & Martin, 1998) or to rapid response learning (Dobbins, Schnyer, Verfaellie, & Schacter, 2004; Logan, 1990).

The repetition priming that is mediated by the PRS has long been thought to be a form of perceptual learning (e.g., Kirsner & Dunn, 1985), so the extensive literature on perceptual learning (e.g., Dosher & Lu, 1999; Fahle & Poggio, 2002) and its neural basis (e.g., Gilbert, Sigman, & Crist, 2001; Petrov, Dosher, & Lu, 2005) might provide answers to these questions. The present results suggest that the category learning literature should closely monitor this debate.

This article reported behavioral dissociations between (A, not A) and (A, B) prototype distortion tasks. In the past, little attention was paid to whether a prototype distortion task was of the (A, not A) or (A, B) variety. We predict that similar results should be seen in other types of category learning tasks, besides prototype distortion, which compare (A, not A) and (A, B) conditions, as long as the A and B categories are characterized by high within-category similarity and the "not A" category members are all perceptually distinct.

The present results add to the growing body of evidence that human category learning recruits multiple memory systems (Ashby & O'Brien, 2005). Different memory systems have different advantages and disadvantages. Since category structures can be of many different types, some memory systems are better adapted than others to learning certain category structures. The present results describe conditions under which the PRS might be critical. Together, all these results suggest that category learning may use many, perhaps all, of the major memory systems that have been hypothesized by memory researchers.

AUTHOR NOTE

This research was supported in part by Public Health Service Grant MH3760-2. We thank David Smith for his helpful comments and suggestions. Correspondence concerning this article should be addressed to F. G. Ashby, Department of Psychology, University of California, Santa Barbara, CA 93106 (e-mail: ashby@psych.ucsb.edu).

REFERENCES

Aizenstein, H. J., MacDonald, A. W., Stenger, V. A., Nebes, R. D., Larson, J. K., Ursu, S., & Carter, C. S. (2000). Complementary category learning systems identified using event-related functional MRI. *Journal of Cognitive Neuroscience*, **12**, 977-987.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, **105**, 442-481.

Ashby, F. G., & Casale, M. B. (2003). The cognitive neuroscience of implicit category learning. In L. Jiménez (Ed.), *Attention and implicit learning* (pp. 109-141). Amsterdam: John Benjamins.

Ashby, F. G., Ell, S. W., & Waldron, E. M. (2003). Procedural learning in perceptual categorization. *Memory & Cognition*, **31**, 1114-1125.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 33-53.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, **37**, 372-400.

Ashby, F. G., & O'Brien, J. B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Sciences*, **2**, 83-89.

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics*, **61**, 1178-1199.

Biederman, I., & Cooper, E. E. (1992). Size invariance in visual object priming. *Journal of Experimental Psychology: Human Perception & Performance*, **18**, 121-133.

Biederman, I., & Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception & Performance*, **19**, 1162-1182.

Brooks, L. (1978). Nonanalytic concept formation and memory for instances. In E. [H.] Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 169-215). Hillsdale, NJ: Erlbaum.

Cave, C. B. (1997). Very long-lasting priming in picture naming. *Psychological Science*, **8**, 322-325.

Cave, C. B., Bost, P. R., & Cobb, R. E. (1996). Effects of color and pattern on implicit and explicit picture memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **22**, 639-653.

Cooper, L. A., Schacter, D. L., Ballesteros, S., & Moore, C. (1992). Priming and recognition of transformed three-dimensional objects: Effects of size and reflection. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 43-57.

DeCaro, M. S., Thomas, R. D., & Beilock, S. L. (2008). Individual differences in category learning: Sometimes less working memory capacity is better than more. *Cognition*, **107**, 284-294.

Demb, J. B., Desmond, J. E., Wagner, A. D., Vaidya, C. J., Glover, G. H., & Gabrieli, J. D. E. (1995). Semantic encoding and retrieval in the left inferior prefrontal cortex: A functional MRI study of task difficulty and process specificity. *Journal of Neuroscience*, **15**, 5870-5878.

Dobbins, I. G., Schnyer, D. M., Verfaellie, M., & Schacter, D. L. (2004). Cortical activity reductions during repetition priming can result from rapid response learning. *Nature*, **428**, 316-319.

Dosher, B., & Lu, Z. L. (1999). Mechanisms of perceptual learning. *Vision Research*, **39**, 3197-3221.

Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology*, **18**, 500-549.

Estes, W. K. (1994). *Classification and cognition*. New York: Oxford University Press.

Fahle, M., & Poggio, T. (Eds.) (2002). *Perceptual learning*. Cambridge, MA: MIT Press.

Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (2002). *Cognitive neuroscience: The biology of the mind* (2nd ed.). New York: Norton.

Gilbert, C. D., Sigman, M., & Crist, R. E. (2001). The neural basis of perceptual learning. *Neuron*, **31**, 681-697.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, **93**, 411-428.

Homa, D., Rhoads, D., & Chambliss, D. (1979). Evolution of concep-

tual structure. *Journal of Experimental Psychology: Human Learning & Memory*, **5**, 11-23.

HOMA, D., STERLING, S., & TREPEL, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning & Memory*, **7**, 418-439.

HOPKINS, R. O., MYERS, C. E., SHOHAMY, D., GROSSMAN, S., & GLUCK, M. (2004). Impaired probabilistic category learning in hypoxic subjects with hippocampal damage. *Neuropsychologia*, **42**, 524-535.

JACOBY, L. L., & HAYMAN, C. A. G. (1987). Specific visual transfer in word identification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **13**, 456-463.

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

KÉRI, S., KELEMEN, O., BENEDEK, G., & JANKA, Z. (2001). Intact prototype learning in schizophrenia. *Schizophrenia Research*, **52**, 261-264.

KIRSNER, K., & DUNN, J. [C.] (1985). The perceptual record: A common factor in repetition priming and attribute retention? In M. I. Posner & O. S. M. Marin (Eds.), *Attention and performance XI* (pp. 547-566). Hillsdale, NJ: Erlbaum.

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

KNOWLTON, B. J., SQUIRE, L. R., & GLUCK, M. A. (1994). Probabilistic classification learning in amnesia. *Learning & Memory*, **1**, 106-120.

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.

KOUTSTAAL, W., WAGNER, A. D., ROTTE, M., MARIL, A., BUCKNER, R. L., & SCHACTER, D. L. (2001). Perceptual specificity in visual object priming: Functional magnetic resonance imaging evidence for a laterality difference in fusiform cortex. *Neuropsychologia*, **39**, 184-199.

LAMBERTS, K. (2000). Information-accumulation theory of speeded categorization. *Psychological Review*, **107**, 227-260.

LOGAN, G. D. (1990). Repetition priming and automaticity: Common underlying mechanisms? *Cognitive Psychology*, **22**, 1-35.

MADDOX, W. T., ASHBY, F. G., ING, A. D., & PICKERING, A. D. (2004). Disrupting feedback processing interferes with rule-based but not information-integration category learning. *Memory & Cognition*, **32**, 582-591.

MADDOX, W. T., BOHIL, C. J., & ING, A. D. (2004). Evidence for a procedural learning-based system in perceptual category learning. *Psychonomic Bulletin & Review*, **11**, 945-952.

MEDIN, D. L., & SCHAFFER, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207-238.

NOSOFSKY, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39-57.

NOSOFSKY, R. M., & JOHANSEN, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, **7**, 375-402.

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.

ODLYZKO, A. M., & SLOANE, N. J. A. (1979). New bounds on the number of unit spheres that can touch a unit sphere in *n* dimensions. *Journal of Combinatorial Theory*, **26**, 210-214.

PETROV, A. A., DOSHER, B. A., & LU, Z.-L. (2005). The dynamics of perceptual learning: An incremental reweighting model. *Psychological Review*, **112**, 715-743.

POSNER, M. I., GOLDSMITH, R., & WELTON, K. E., JR. (1967). Perceived distance and the classification of distorted patterns. *Journal of Experimental Psychology*, **73**, 28-38.

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

POSNER, M. I., & KEELE, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, **83**, 304-308.

RAICHLE, M. E., FIEZ, J. A., VIDEEN, T. O., MACLEOD, A.-M. K., PARDO, J. V., FOX, P. T., & PETERSEN, S. E. (1994). Practice-related changes in human brain functional anatomy during nonmotor learning. *Cerebral Cortex*, **4**, 8-26.

REBER, P. J., & SQUIRE, L. R. (1999). Intact learning of artificial grammars and intact category learning by patients with Parkinson's disease. *Behavioral Neuroscience*, **113**, 235-242.

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

REBER, P. J., STARK, C. E. L., & SQUIRE, L. R. (1998b). Cortical areas supporting category learning identified using functional MRI. *Proceedings of the National Academy of Sciences*, **95**, 747-750.

REED, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, **3**, 382-407.

ROEDIGER, H. L., III, & BLAXTON, T. A. (1987). Effects of varying modality, surface features, and retention interval on priming in word-fragment completion. *Memory & Cognition*, **15**, 379-388.

ROEDIGER, H. L., III, & SRINIVAS, K. (1993). Specificity of operations in perceptual priming. In P. Graf & M. E. J. Masson (Eds.), *Implicit memory: New directions in cognition, development, and neuropsychology* (pp. 17-48). Hillsdale, NJ: Erlbaum.

ROSCH, E. H. (1973). Natural categories. *Cognitive Psychology*, **4**, 328-350.

ROSCH, E. [H.] (1975). Cognitive reference points. *Cognitive Psychology*, **7**, 532-547.

SCHACTER, D. L. (1990). Perceptual representation systems and implicit memory: Toward a resolution of the multiple memory systems debate. *Annals of the New York Academy of Sciences*, **608**, 543-571.

SCHACTER, D. L., & BUCKNER, R. L. (1998). Priming and the brain. *Neuron*, **20**, 185-195.

SCHACTER, D. L., WIG, G. S., & STEVENS, W. D. (2007). Reductions in cortical activity during priming. *Current Opinion in Neurobiology*, **17**, 171-176.

SEAMON, J. G., GANOR-STERN, D., CROWLEY, M. J., WILSON, S. M., WEBER, W. J., O'ROURKE, C. M., & MAHONEY, J. K. (1997). A mere exposure effect for transformed three-dimensional objects: Effects of reflection, size, or color changes on affect and recognition. *Memory & Cognition*, **25**, 367-374.

SEGER, C. A., POLDRACK, R. A., PRABHAKARAN, V., ZHAO, M., GLOVER, G. H., & GABRIELI, J. D. E. (2000). Hemispheric asymmetries and individual differences in visual concept learning as measured by functional MRI. *Neuropsychologia*, **38**, 1316-1324.

SHEPARD, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, **237**, 1317-1323.

SHIN, H. J., & NOSOFSKY, R. M. (1992). Similarity-scaling studies of dot-pattern classification and recognition. *Journal of Experimental Psychology: General*, **121**, 278-304.

SINHA, R. R. (1999). Neuropsychological substrates of category learning. *Dissertation Abstracts International*, **60**(5-B), 2381. (UMI No. AEH9932480)

SMITH, J. D., & MINDA, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1411-1436.

SMITH, J. D., & MINDA, J. P. (2001). Journey to the center of the category: The dissociation in amnesia between categorization and recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 984-1002.

SMITH, J. D., & MINDA, J. P. (2002). Distinguishing prototype-based and exemplar-based processes in category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 800-811.

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.

WAGNER, A. D., DESMOND, J. E., DEMB, J. B., GLOVER, G. H., & GABRIELI, J. D. E. (1997). Semantic repetition priming for verbal and pictorial knowledge: A functional MRI study of left inferior prefrontal cortex. *Journal of Cognitive Neuroscience*, **9**, 714-726.

WALDRON, E. M., & ASHBY, F. G. (2001). The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. *Psychonomic Bulletin & Review*, **8**, 168-176.

WIGGS, C. L., & MARTIN, A. (1998). Properties and mechanisms of perceptual priming. *Current Opinion in Neurobiology*, **8**, 227-233.

ZAKI, S. R., & NOSOFSKY, R. M. (2004). False prototype enhancement effects in dot pattern classification. *Memory & Cognition*, **32**, 390-398.

ZAKI, S. R., NOSOFSKY, R. M., JESSUP, N. M., & UNVERSAGT, 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.

ZEITHAMOVA, D., & MADDOX, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition*, **34**, 387-398.

NOTES

1. The exact values used to group the Category A stimuli in (A, B) tasks were $(D_{xA} - D_{xB} + D_{AB})/2$, where D_{AB} is the distance between the prototypes of Categories A and B. Adding the D_{AB} constant and dividing by 2 ensures that the scales for the (A, not A) and (A, B) conditions are equal. For example, suppose that $D_{AB} = 100$, and consider a stimulus on the line connecting the two prototypes that is 30 units from the A prototype. Then, in the (A, not A) conditions, this stimulus would be placed in the bin marked 30 (since $D_{xA} = 30$). In the (A, B) condition, this same stimulus would also be placed in the bin marked 30 [since $(D_{xA} - D_{xB} + D_{AB})/2 = (30 - 70 + 100)/2 = 30$].

2. This is because the category exemplars are created by distorting the prototype by adding independent and identically distributed noise to each stimulus dimension (e.g., Ashby & Gott, 1988).

3. These likelihood predictions follow because we created the Category A members by distorting the horizontal and vertical locations of each dot in the prototype by adding samples from independent and identically distributed normal distributions (with mean zero).

4. An exemplar-based model that would be more consistent with a pure PRS strategy would replace the summed similarity in Equation 4 with the summed similarity of the presented stimulus to all previously seen exemplars, regardless of their category membership (i.e., rather than just to all previously seen Category A exemplars). This is because the PRS includes no record of the category membership of previously viewed stimuli. Note that this sum should be roughly equal for an atypical Category A exemplar that was highly similar to some previously seen Category A member and for a "not A" exemplar that was highly similar to some previously seen "not A" member.

5. One might choose an optimal solution, such as the correspondence that minimizes the Euclidean distance between the two patterns. There are two problems with this approach, however. First, there is no empirical evidence that such a choice gives a better account of perceived similarity than other choices, and second, there are many different ways to define the optimal solution (e.g., minimize bending required to bring the two patterns into alignment).

6. By this, we mean we did nothing. Therefore, for Category A patterns we assumed that each dot corresponded to the dot in the prototype pattern from which it was distorted. For "not A" patterns, we randomly labeled each dot.

7. Note that this highly nonintuitive property of high-dimensional spaces reduces the importance of the dot correspondence problem. Especially with "not A" category members, many stimuli will be approximately equidistant from each other, regardless of which dot assignment choice is made.

(Manuscript received October 29, 2007;
revision accepted for publication February 26, 2008.)