Predicting Reasoning From Memory

Evan Heit
University of California, Merced

Brett K. Hayes
University of New South Wales

In an effort to assess the relations between reasoning and memory, in 8 experiments, the authors examined how well responses on an inductive reasoning task are predicted from responses on a recognition memory task for the same picture stimuli. Across several experimental manipulations, such as varying study time, presentation frequency, and the presence of stimuli from other categories, there was a high correlation between reasoning and memory responses (average $r = .87$), and these manipulations showed similar effects on the 2 tasks. The results point to common mechanisms underlying inductive reasoning and recognition memory abilities. A mathematical model, GEN-EX (generalization from examples), derived from exemplar models of categorization, is presented, which predicts both reasoning and memory responses from pairwise similarities among the stimuli, allowing for additional influences of subtyping and deterministic responding.

Keywords: inductive reasoning, recognition memory, similarity, mathematical modeling

Inductive inference involves extending knowledge from known instances to novel instances and is a central component of human learning and reasoning. Such reasoning leads to conclusions that are probable rather than certain and is therefore thought to be critical to people's everyday encounters with an uncertain world. For example, after learning that lions, horses, and otters use dopamine as a neurotransmitter, one might make a reasonably confident (but by no means certain) inference that this property is shared by other mammals. Many believe that inductive reasoning is the most important component of human thought because "inductive inference is the only process . . . by which new knowledge comes into the world" (Fisher, 1951, p. 7).

In 2 decades of research, a number of general principles that people follow when making inductive inferences has been identified (see Hayes, Heit, & Swendsen, 2010; Sloman & Lagnado, 2005, for reviews). The cognitive mechanisms underlying induction and the relation between these mechanisms and those underlying other cognitive phenomena such as categorization and memory, however, remain matters of intense debate (cf. Hayes & Heit, 2004).

This article addresses two shortcomings in current process theories of induction. First, most current models of induction (e.g., Kemp & Tenenbaum, 2009; Osherson et al., 1990; Sloman, 1993) have been framed to explain how novel properties are generalized from one or more categories of objects to other categories (as in the neurotransmitter example above). Although this kind of reasoning is important, it is also essential to explain how people generalize novel properties involving individual exemplars. Knowing the general characteristics of dogs does not necessarily help you to make an inference about whether a neighbor's dog will be fierce or friendly. To make such an inference, people may need to go beyond category-level information and consider the specific features of the dog, like size and propensity to growl on approach, and to compare these features with similar dogs encountered in the past. Existing models do not address the role of similarity between specific instances in inductive projection. One of the aims of this article is to redress this imbalance by outlining a model of induction that specifies how similarity between individual exemplars drives induction.

A second serious shortcoming of existing induction models is that they treat inductive reasoning as disconnected from other cognitive processes. In particular, previous models of induction have not addressed the role of memory. Even a cursory consideration of inductive reasoning, however, suggests that memory should play a central role. Being able to remember the similarities (and differences) between lions and otters seems central to explaining how a property shared by these categories will be generalized. Being able to retrieve memories of dogs that are similar to the one that lives next door is useful when making predictions about that dog's behavior. Despite the central role of memory in induction, there have been few attempts to examine the specific connections between the processes involved in each task. Instead each has been most often studied with its own experimental paradigms, addressing different questions and resulting in induction (Hayes et al., 2010; Sloman & Lagnado, 2005) and memory phenomena (Diana, Reder, Arndt, & Park, 2006; Wixted, 2007) being addressed by separate theories.

In this article, we address this problem by focusing on the links between inductive reasoning and recognition memory. Despite the apparent differences between these tasks, we argue that they share some key underlying processes and that in many contexts there will be a close relation between inductive inferences and memory for a common stimulus set. We also outline a computational model that embodies this close relation and that can account for both inductive and recognition judgments about the same stimuli.

The Case for Common Processes in Induction and Recognition

At a general level, induction about specific objects and object recognition, like many other perceptual and cognitive tasks, involves the generalization of existing knowledge about familiar stimuli and their properties to novel stimuli (Shepard, 1987). A more specific point of overlap is the central role accorded in each task to similarity between familiar and novel exemplars. In recognition, the probability that an item is recognized as "old" is a positive function of its similarity to previously studied items (Jones & Heit, 1993). As noted above, the role of exemplar similarity in induction has not been examined in detail, but it seems safe to assume that the probability that a novel item is judged to have a property depends, in part, on its similarity to known instances that have that property.

The proposed overlap between recognition and induction goes beyond the level of task description; existing models of both recognition (Hintzman, 1988; Ratcliff, 1990) and induction (Osherson et al., 1990; Sloman, 1993) view similarity computation as a core process that determines performance. Despite this apparent overlap in generalization processes, most models of recognition memory have not addressed reasoning, and likewise, previous models of inductive reasoning have not addressed memory (Heit & Hayes, 2005).

Of course both recognition and induction may involve other processes besides an assessment of old–new similarity. Dual-process models of memory, for example, posit that recognition is driven by a fast, automatic response based on stimulus familiarity and a slower, deliberative recollection judgment that involves episodic components like where and when the stimulus was seen (e.g., Wixted, 2007; Yonelinas, 2002). Likewise, some models of induction suggest that more complex principles like causal relations between an inductive base and a target can supplement or even supplant similarity as a basis for generalization (Kemp & Tenenbaum, 2009; Lee & Holyoak, 2008; Medin, Coley, Storms, & Hayes, 2003; Rehder & Burnett, 2005). We do not dispute that recognition and induction can involve more than a consideration of similarity between familiar and novel instances (an issue to which we return in the General Discussion). The key point is that even such multiple-process models acknowledge that assessing the similarity of a new item to familiar items is a core process that drives both recognition and inductive judgments. Moreover, in our own work, we will look for evidence of other processes beyond similarity.

Our argument for close links between induction involving specific exemplars and recognition is also motivated by previous work that has discovered strong empirical and theoretical links between inductive reasoning and categorization (Rehder & Burnett, 2005;

Sloutsky & Fisher, 2004) as well as categorization and recognition memory (Estes, 1994; Nosofsky, 1988a).

Rehder and Burnett (2005), for example, found a strong empirical relation between induction and categorization, with correlations across items of .70 to .99 between inferences about whether an exemplar possessed a novel property and category membership judgments for the same items. Sloutsky and Fisher (2004) also observed a strong correlation across items between children's category membership and judgments about the generalization of novel properties for both naturalistic and artificial stimuli.

With regard to the relation between categorization and recognition, exemplar models of categorization, such as the highly successful generalized context model (GCM; Nosofsky, 1986), which embodies Shepard's (1987) proposed universal law of generalization, have made the case for a systematic relation between performance on these two tasks. In support of this argument Nosofsky (1988a) showed that that old–new recognition judgments and categorization judgments of the same items can be explained by assuming that both are subserved by a common memory trace for exemplars but that different decision rules are applied in each task. In a similar vein, Lamberts (2002) has shown that the time course of recognition and classification responses to the same items can be explained by models that assume a common representation for each judgment.

These various theoretical arguments and lines of evidence all suggest that the specific similarity between exemplars is a common core process in recognition and induction and that it should therefore be possible to uncover reliable empirical relations between recognition and induction judgments. In the next section, we outline a paradigm that allows us to examine the relation between recognition and induction and develop a number of specific predictions about how this relation will behave across a range of encoding and test manipulations.

Comparing Recognition and Induction

To examine inductive reasoning about specific instances and to explore the relation between recognition and induction, we developed a new experimental paradigm that makes reasoning and memory tasks as comparable as possible. In all of the experiments reported, people were asked to either make recognition judgments about a set of pictures they had studied or make property inferences about the same set. We examined whether the overgeneralization errors that people make in visual recognition predict the pattern of generalization that other people show in inductive reasoning.

When memory and reasoning tasks differ only in the nature of judgments being made, we make two general predictions. First, we expect that induction and recognition instructions will lead to a systematic difference between the tasks in the breadth of generalization to novel test instances. Recognition instructions emphasize that a positive response to a test item should only be made if that item has been studied. Property induction instructions, on the other hand, explicitly invite the participant to go beyond the information provided in inductive premises and project a property to novel items (e.g., Osherson et al., 1990; Sloman, 1993). These instructions should lead to a higher rate of positive responding to novel items under induction conditions than recognition conditions.

Second, if both induction and recognition share a common underlying process, there should be a reasonably close correspondence between them in the pattern of positive responding for individual test items. Because exemplar similarity is expected to influence responses on both tasks, items that are more likely to be identified as old should generally be judged as strong candidates for property inference.

A further aim of these experiments was to examine the robustness of the relation between memory and reasoning across a variety of task manipulations. Conventional approaches to visual recognition and induction assume that each is driven by different kinds of processes. Recognition, for example, is assumed to be strongly influenced by perceptual similarity between study and test items (Lamberts, 2002), whereas inductive reasoning also involves more complex semantic or causal relations between premises and conclusions (Kemp & Tenenbaum, 2009; Medin et al., 2003). According to such approaches, a range of task factors might selectively affect performance on one kind of task without affecting the other. Manipulation of the frequency of presentation of individual study items, for example, is known to affect recognition responding (e.g., Flexser & Bower, 1974; Hintzman, 2001). By contrast, this factor seems less likely to directly impact semantic relations and therefore might be expected to have little or no effect on induction (but see Xu & Tenenbaum, 2007). Going in the other direction, if items from multiple basic categories were all found to have some novel property in common, this should strengthen the generalization of this property to novel members of one of these categories (Osherson et al., 1990) but should have little effect on recognition of old and new category members.

According to our approach, however, any factor that affects the specific similarity between studied items and novel test cases will affect both recognition and induction responses, often in similar ways. In other words, whereas conventional approaches to recognition and induction suggest a range of factors that should show empirical dissociations across these tasks, we predict that these factors should often have parallel effects on memory and reasoning. These contrasting predictions were investigated in a series of studies in which the same manipulations were applied to comparable recognition and induction tasks. These included study phase encoding time (Experiments 1A and 1B), item frequency during study (Experiments 2A and 2B), and the presence of items from multiple basic categories during study (Experiment 3) or test phases (Experiment 4). In each case, we expected that these manipulations would have similar effects on the computation of old–new similarity and hence should have comparable effects on the probability of making a positive response in induction and recognition tests.

GEN-EX (Generalization From Examples): An Exemplar-Based Model of Inductive Reasoning and Recognition Memory

An important goal of this work was to examine whether reasoning and memory performance could be accommodated within a single computational model. The core assumption of our model of reasoning and memory is adapted from the generalized context (GCM) family of categorization models (e.g., Nosofsky, 1986, 1988a). We assume that the tendency to make a positive ("yes") response to a test stimulus in induction and recognition is a positive function of the total similarity between that stimulus and all studied items. Exemplar models have been successful in accounting for patterns of categorization and recognition of the same stimulus sets (e.g., Nosofsky, 1986; Shin & Nosofsky, 1992) but have only rarely been applied to inductive reasoning data (e.g., Estes, 1994; Heit, 1992). An important advantage of exemplar models is that they can account for empirical dissociations between tasks without assuming multiple cognitive systems. Nosofsky and Zaki (1998), for example, showed that dissociations between recognition and categorization performance in amnesic patients and normal controls could be explained with a single, exemplar-based model. They allowed the shape of the generalization gradient to vary between groups and a response criterion parameter to vary from the categorization task to the recognition task. Note, however, that Nosofsky (1991) found that only changing the response criterion was insufficient to account for all differences between categorization and recognition and concluded that attention shifted between tasks, as well.

In our own work, comparing induction and recognition, we allow for the possibility of different response criteria, but we also investigate the possibility that the shape of the generalization gradient will be different for induction and recognition. The key prediction is that induction would involve broader generalization to novel items, whereas recognition would be more sensitive to exact matches between studied and test items. After all, the nature of the task of induction is to generalize to unseen instances (e.g., Heit, 2007), whereas the goal of the recognition task is to reject unseen instances. Relatedly, Lamberts (1994) compared two kinds of categorization tasks, involving categorization of faces as brothers or cousins, and using a model-based analysis, found a broader generalization gradient for the cousin categorization task.

The basic version of GEN-EX (so named because it generalizes from examples) is embodied by two equations. Equation 1 shows the familiarity rule. The familiarity, fam, of each test stimulus equals its summed similarity to n studied items. Similarity is assumed to be an negative exponential function of distance, dist, between the test and the study items, calculated according to the standard Euclidean formula. The free parameter c reflects specificity of responding to test items; lower values of c correspond to broader generalization, whereas higher values correspond to narrower generalization gradients. Put another way, the c parameter reflects "overall discriminability in the psychological space" (Nosofsky, 1986, p. 41), with higher values indicating a greater level of discrimination.

$$\text{fam(test)} = \sum_{i=1}^{n} \exp[-c\, \text{dist}\,(\text{test, stud}y_i)] \qquad (1)$$

$$\text{resp(test)} = \frac{\text{fam(test)}}{\text{fam(test)} + \beta} \qquad (2)$$

The response rule is shown in Equation 2. Essentially, the probability of a positive response, *resp*, is a monotonic function of a test item's familiarity. The response rule has a single scaling parameter, β. A lower value of β corresponds to a greater overall tendency to respond positively. Nosofsky and Zaki (1998) used such a scaling parameter to help account for differences between categorization and recognition.

Note that although GEN-EX is closely related to Nosofsky's GCM model, the two models are not isomorphic. GCM was developed to explain object classification in which a novel probe is assigned to one of a number of possible category alternatives. GCM is not a model of reasoning and was never intended to explain the generalization of object properties between exemplars. One critical difference between classification and property induction (and hence between GCM and GEN-EX) is that classification responses are based on the ratio of similarities between a given probe and a range of possible categories. In contrast, GEN-EX assumes that property induction between instances is based only on the similarities between instances within a given category. Similar assumptions about the primacy of within-category similarity have been made in previous quantitative models of induction (e.g., Yamauchi & Markman, 1998). Moreover, in applying GEN-EX to experimental data, we incorporated two kinds of deterministic responding, based on subtyping and item recollection, which are not present in GCM.

The implementation of GEN-EX and its predictions with regard to test responses on comparable induction and recognition tests are laid out in more detail in the modeling section following Experiment 1A. The central predictions, however, were that positive responses under recognition and induction instructions should be well predicted by GEN-EX with the same old–new similarity values for test items but that the c parameter should reflect broader levels of generalization under induction, as compared with recognition conditions.

Experiment 1A

For this experiment we developed a new experimental paradigm that makes inductive reasoning and recognition tasks as comparable as possible (see Heit & Hayes, 2008, for a preliminary report). In brief, in the induction condition, people were asked to learn about instances from a single category (large dogs) that shared a novel property (e.g., "has beta cells inside"), whereas those in the recognition condition were asked to memorize the same instances. Both groups were then shown a common test set that contained both old instances and a range of new instances that varied in similarity to old instances (i.e., new large dogs that functioned as lures as well as new medium-sized and small dogs). In the recognition condition, people responded "yes" if they thought a test item had been presented during the study phase. In the induction condition, people responded "yes" if they thought a test item had the target property.

Overall, we expected that people doing induction would generalize study information more broadly than those doing recognition, leading to higher false alarm rates and lower d' sensitivity. Nevertheless, because recognition and induction both rely on exemplar similarity, it was expected that there would be a high positive correlation between the probabilities of responding "yes" to specific test items across the two tasks.

Method

Participants. Sixty-three students were recruited individually in quiet, public places, such as the library, on the University of California, Merced campus. Participants were randomly assigned to two conditions: memory ($n = 31$) and reasoning ($n = 32$).[1]

Materials. The stimuli were color photographs of dogs, 280 pixels square, adapted from a compendium of dog breeds (American Kennel Club, 2006) and other Internet sources. Each photograph showed a dog in a canonical left-facing side view. The same stimulus set was used for both conditions. The study list consisted of 10 pictures of large dogs. The test list consisted of 45 pictures of dogs. There were 10 old items (the large dogs originally studied), 15 lure items (other large dogs, not previously studied), and 20 additional, new items (10 small dogs and 10 medium dogs).[2]

Procedure. The experiment was run with a program on a laptop computer. In the memory condition, participants were instructed to memorize the initial set of pictures, for a subsequent recognition test. They were shown the 10 pictures on the study list, in a different random order for each participant. Each 10 cm^2 picture was presented for 2 s, with a 0.5 s interstimulus interval during which the screen was blank. There was a 60 s unfilled retention interval before the test phase. Participants were instructed to judge whether they had seen each test picture by clicking their mouse on either a *yes* or *no* button on the computer screen. Each test item remained on the screen until a response was made, and there was no time limit on responding. After each response, there was a 0.5 s interstimulus interval during which the screen was blank. During the test phase, the 45 test pictures were shown sequentially, in a different random order for each participant.

The reasoning condition was like the memory condition, except for the following. Before the study phase, participants were told they would study a set of animals with "beta cells" in their blood, for a subsequent test on whether various old and new items have this property. During the test phase, participants were asked to judge whether each animal has "beta cells."

Results

Probability of responding "yes." The probability of responding "yes" to test items under recognition and induction conditions is shown in Table 1. In the memory condition, recognition performance was good, with a relatively high hit rate on old items (large dogs) and a false alarm rate of .15 on new items. The false alarm rate was slightly higher on pictures of medium dogs than pictures of small dogs. For the lure items (large dogs not studied), the average false alarm rate was .30.

---

[1] Experiments 1A and 1B included a second reasoning condition, with a behavioral property, "performs behavior X," rather than an anatomical property. Although we had surmised that manipulating type of property would make a difference (Heit & Rubinstein, 1994), in fact, the behavioral and anatomical reasoning conditions did not differ significantly on any measures. Hence, we report only the anatomical reasoning conditions here. We suspect that finding evidence for two different patterns of reasoning would require a stronger manipulation of property type and, as Heit and Rubinstein used, a choice of materials that unconfounds the usual correlation between anatomy and behavior.

[2] We note that unlike some experimental studies of memory, in the present studies, there was a higher proportion of new items than old items in the test list. This may have led participants to be conservative in their responding (i.e., tending to say "no"; Heit, Brockdorff, & Lamberts, 2003). However this aspect of the design was crucial for a detailed examination of changes in test responding as a function of the similarity of the novel test items and for the rigorous testing of the quantitative models.

Table 1
Experiment 1A. Results (Proportion of "Yes" Responses and d′*) and Model Predictions*

Result and prediction	Old	New small	New medium	All new	Lure	d′ (Old–new)	d′ (Old–lure)
Empirical results							
Memory	.68	.13	.17	.15	.30	1.50	0.98
Reasoning	.82	.41	.49	.45	.68	1.03	0.44
Model predictions							
Memory	.68	.09	.18	.14	.30	1.56	0.98
Reasoning	.82	.40	.50	.45	.68	1.03	0.43

Compared with the recognition condition, participants in the induction condition were more likely to give positive responses. On old items, they inferred beta cells .82 of the time, and on new items, .45 of the time. As in the memory condition, there were more positive responses to medium dogs than to small dogs. For the lure items, the rate of positive response was high: .68. Compared with recognition, in the induction condition there was a high rate of generalization, with participants particularly likely to extend the property to the lure items.

These observations were confirmed with a series of analyses of variance (ANOVAs) assessing hit rates, false alarm rates, and d′. These ANOVAs compared the recognition and induction conditions, as a between-subjects variable, separately for various stimulus types. The probability of responding positively to old items was higher in the induction condition than in the recognition condition, $F(1, 61) = 9.07$, $p < .01$, partial $\eta^2 = .13$. Positive responses were more common for new medium than for new small dogs, $F(1, 61) = 8.20$, $p < .01$, partial $\eta^2 = .12$, but this difference did not interact with task. Hence, in subsequent analyses, responses to small and medium dogs were collapsed into a set of new test items. The probability of positive responding to both new and lure items was higher for induction than for recognition, $F(1, 61) = 17.09$, $p < .001$, partial $\eta^2 = .22$, and $F(1, 61) = 34.06$, $p < .001$, partial $\eta^2 = .36$, respectively.

To further examine patterns of generalization across the memory and reasoning conditions, a d′ measure of sensitivity was calculated for each participant, with individual hit rates and false alarm rates for new (small and medium) dogs and lure items, respectively. Individual hit and false alarm rates of zero or one were corrected with the procedure outlined by Macmillan and Kaplan (1985), whereby a value of 1/(2N) was added to rates of 0 and subtracted from rates of 1, with N equal to the number of test trials for a given type of test stimulus. This meant that a defined value of d′ could be calculated for all participants. This correction method was used in all subsequent experiments.

The mean sensitivity values for each condition are given in Table 1. Sensitivity in the discrimination between old and new items tended to be higher for recognition than induction, but this difference was not reliable, $F(1, 61) = 3.00$, $p = .09$. Sensitivity in the discrimination between old and lure items was significantly higher for recognition than induction, $F(1, 61) = 8.26$, $p < .01$, partial $\eta^2 = 0.12$.

Relation between reasoning and memory. The proportion of "yes" responses for each of the 45 test items was averaged across participants within each of the two experimental conditions, and the correlation between responses in different conditions was computed. The correlation between the memory condition and the

reasoning condition was .83, ($p < .001$). In other words, memory was a very good predictor of reasoning. This relation is illustrated in Figure 1, showing a scatter plot of memory responses versus reasoning, for the 10 old items; 15 lure items; 10 new, medium dogs; and 10 new, small dogs. (Some of the data points overlap so closely that all 45 data points may not be discriminable.) Note that there is a greater level of generalization for reasoning than for memory.

Modeling

To model recognition and induction judgments with GEN-EX, we first needed to collect pairwise similarity ratings between pairs of study and test items. (Note that we previously applied an earlier version of GEN-EX to this data set in Heit & Hayes, 2008, using simulated similarity relations rather than real data). There were 10 study items, and 45 test items, giving a total of 450 pairs. These pairs were presented on a computer screen in random order to a
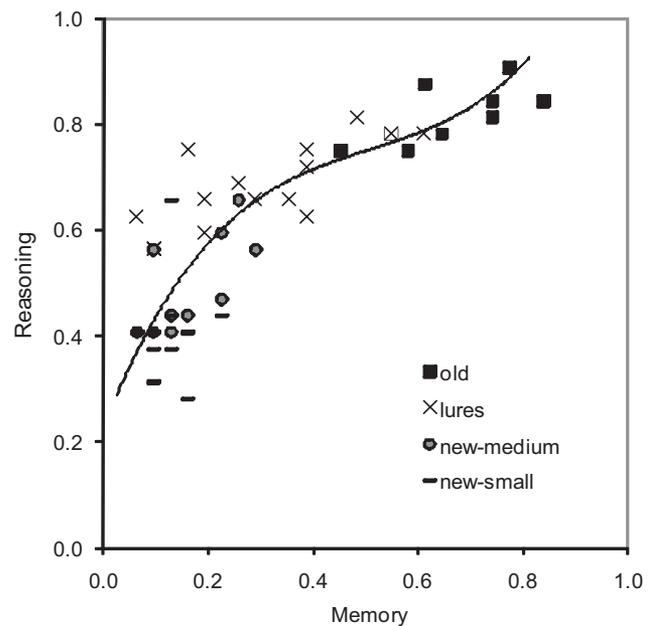

Figure 1. Scatter plots for Experiment 1A, showing memory and reasoning responses across stimuli, with trend line of GEN-EX $s + d$ model predictions. Note that some data points overlap so closely that all 45 data points may not be discriminable. GEN-EX = generalization from examples.

group of 40 participants who did not take part in the main experiment. Each pair was presented for a minimum of 3 s, after which participants were asked to rate the similarity between the items on a 7-point scale (1 = *not similar*; 7 = *highly similar*) while the pictures remained on the screen. The left–right position within a pair of pictures was determined randomly on each trial. Participants were given periodic breaks to minimize fatigue. An estimate of the similarity between each study and test item was then obtained by averaging the relevant ratings for each study–test pair across participants. These estimates were normalized to the range of 0 to 1 by subtracting 1 then dividing by 6.

It was assumed that similarity, sim, would be a negative exponential function of psychological distance, dist, (e.g., Nosofsky, 1986), as illustrated in Equation 3.

$$\text{sim}(x, y) = \exp[-c\, \text{dist}(x, y)] \qquad (3)$$

There were three c parameters. One was used for converting similarity ratings to distances, as in Equation 3. That is, Equation 3 was solved for distance as a logarithmic function of similarity, to calculate distances as a function of an estimated c parameter. The other two were used for converting distances back to similarity, as in Equation 1, when calculating familiarity for recognition and for induction. In addition, there were two β parameters, serving as scaling parameters for recognition and induction.

We developed multiple versions of GEN-EX, to take account of other possible response patterns not predicted by item similarity. Loosely speaking, these patterns could be thought of as deterministic or rule-based responding. One pattern would be to form a subcategory corresponding to the studied items. Here, the subcategory would be large dogs. A participant responding according to subcategories would respond positively to large dogs, whether old or lures, and respond negatively to small and medium dogs (see Bott & Murphy, 2007, and Hayes et al., 2002, for related ideas.) Informally, such a participant would have induced that large dogs have the characteristic of interest, whether that is having been presented or possessing some property. GEN-EX s was designed to capture this pattern of responding.

Another possible pattern would be to respond deterministically, depending on whether the item was actually studied, as if memory was perfect. A participant responding deterministically would respond positively to old items and negatively to new items and would not be willing to generalize. In terms of memory models, this pattern is equivalent to recollection (Yonelinas, 2002). GEN-EX d was developed to examine this kind of pattern.

The most complete version of the model, GEN-EX $s + d$, had free parameters corresponding to both of these response patterns as well as responding according to item similarity, as show in Equation 4.

$$\text{resp(test)} = s \cdot \text{large(test)} + d \cdot \text{old(test)}$$
$$+ (1 - s - d)\frac{\text{fam(test)}}{\text{fam(test)} + \beta} \qquad (4)$$

Here, s represents the probability of responding according to subcategories (responding positively whether the test item is a large dog); the function large takes on a value of 1 for a large dog and 0 otherwise. Likewise, d represents the probability of responding deterministically according to whether the item is old; the

function old takes on a value of 1 for a studied item and 0 otherwise. Distinct s and d parameters were estimated for recognition and induction, to allow for different response patterns for the two tasks. For example, it seemed possible that induction might entail more subcategory based responding and that recognition would entail more deterministic responding. There were also three restricted models. In GEN-EX s, the d parameter was set to zero, so there was no deterministic responding. In GEN-EX d, the s parameter was set to zero, so there was no subcategory responding. Finally, in GEN-EX null, both the s and d parameters were set to zero; this is the original GEN-EX model.

The four GEN-EX models were used to generate predictions for 90 data points, corresponding to the 45 test items for memory and for inductive reasoning, respectively. The free parameters were estimated with the solver function in Microsoft Excel, minimizing the root-mean-square error (*RMSE*; the square root of the mean squared difference between model predictions and response proportions across items) of prediction. The model fitting is summarized in Table 2, including goodness of fit measures for the GEN-EX $s + d$ model, significance tests comparing the $s + d$ model to the s, d, and null models, and estimated parameter values for the $s + d$ model. Although all four versions of the GEN-EX model fit the data reasonably well, the GEN-EX $s + d$ model fit significantly better than each of the three restricted models, after taking account of number of parameters (having a correlation of .94 with the data across 90 data points, using 9 free parameters).[3] Hence, both subcategory responding and deterministic responding components led to significant improvements in the fit of the model.

Table 2 shows the estimated parameter values for GEN-EX $s + d$. Note that as predicted, the c value was considerably higher for memory than for reasoning, reflecting sharper generalization for memory and broader generalization for reasoning. The s and d parameters were estimated to be nonzero but fairly low, so that the greatest overall influence on responses was item similarity rather than subtyping or deterministic recollection. The value of s was estimated to be higher for induction than for recognition, suggesting a greater influence of sub-categories for induction. In contrast, d was estimated to be higher for recognition than for induction, suggesting a greater tendency to respond deterministically for recognition. Finally, the β parameter values are similar for memory and reasoning; there is little evidence for different response scaling between tasks.

Table 1 shows average predictions of the GEN-EX $s + d$ model, for key types of stimuli. The table shows that the main trends in the data have been captured, such as differences between memory and reasoning conditions and differences between old, lure, new medium, and new small items. Likewise, the predicted d' measures are close to the original results. Note that the simulation had actually made 45 predictions each for the memory and reasoning

---

[3] The nested models were compared with the technique of Borowiak (1989). In brief, when Model A is a nonlinear model with a free parameters estimated with a least-squares criterion and B is a restricted version of this model with b free parameters, the likelihood ratio statistic is $\lambda = (\text{RSS}_A/\text{RSS}_B)^{(k/2)}$, where RSS is the residual sum of squares of the model and k is the number of data points to be predicted (here, 20). Borowiak showed that $-2 \ln (\lambda)$ has a χ^2 distribution with $(a - b)$ degrees of freedom. (See Heit, 1998b, 2001, and Rotello & Heit, 1999, 2000, for other applications of this technique.)

Table 2
Summary of Model Fitting for GEN-EX s + d

Measure	Expt. 1A	Expt. 1B	Expt. 1C	Expt. 1D	Expt. 2A Freq. 1	Expt. 2A Freq. 3	Expt. 2B Vers. A	Expt. 2B Vers. B
Data points	90	90	90	45	180		180	
RMSE	.0859	.0993	.0887	.0426	.1127		.1260	
Correlation	.9409	.8825	.9533	.9865	.8749		.9022	
χ^2 over *s* model	7.45*	18.40***	7.86*	108.62***	20.94***		20.50***	
χ^2 over *d* model	6.86*	4.15, *ns*	3.94, *ns*	41.29***	0.00, *ns*		0.00, *ns*	
χ^2 over null model	12.35*	19.67***	11.48*	108.62***	20.94**		20.50**	
c sim	2.18	1.73	2.89	1.70	1.80		2.58	
c rec	3.56	3.65	3.07		4.45	4.76	4.16	4.38
c ind	1.91	1.41	3.31	1.69	1.64	2.64	2.14	3.30
β rec	0.79	1.01	0.42		0.36	0.30	0.30	0.39
β ind	0.66	2.35	0.16	0.69	1.53	0.77	0.79	0.42
s rec	.01	.00	.00		.00	.00	.00	.00
s ind	.11	.10	.08	.03	.00	.00	.02	.00
d rec	.17	.19	.29		.10	.21	.24	.09
d ind	.09	.19	.00	.57	.17	.17	.28	.11

Note. The table shows goodness of fit (root mean squared error and correlation) for the GEN-EX *s + d* model, comparative tests showing improvement in goodness of fit over the *s*, *d*, and null models, and estimated parameter values for the *s + d* model. GEN-EX = generalization from examples model; Expt. = experiment; Freq. = frequency; Vers. = version; *RMSE* = root-mean-square error; sim = similarity; rec = recognition; ind = induction.
* $p < .05$. ** $p < .01$. *** $p < .001$.

conditions, across the 45 test items. Thus, it was possible to calculate the predicted correlation between memory and reasoning, just as the original data had shown a .83 correlation between memory and reasoning across the 45 test items. In the simulations, the average predicted correlation between memory and reasoning was .92. This result was an emergent property of the model; that is, the correlation results were not fitted directly by the model.

In Figure 1, model predictions are shown as a trend line, which was derived by fitting a third-degree polynomial function to the model's predictions on individual items. The purpose of doing so is to assess whether the model predictions fall in the region of the data points for different types of items. Overall, the trend predicted by the model well captures the trend in the data points, although there is some scatter in the data points not predicted by the trend line derived from model predictions. One key trend is that the model predicts a higher rate of positive responding for the reasoning condition than for the memory condition, even for new and lure items, a consequence of the lower *c* parameter for reasoning, leading to higher familiarity values in Equation 1.

Discussion

This experiment examined the relation between recognition and inductive judgments about the same set of visual stimuli under comparable study and test conditions. The main empirical finding was that when procedural differences between these tasks were kept to a minimum, there was a close correspondence between the two kinds of judgments. Judgments about whether a novel property would generalize to a given test item was positively related to the probability of responding "old" to that item in recognition. The strength of this relation supports the view that recognition and induction share some underlying component processes.

There were also some interesting differences between recognition and inductive reasoning. People doing induction were more likely to make positive responses to both familiar and novel test stimuli (and hence showed lower sensitivity) than those doing recognition. Induction instructions promoted a broad generalization of the novel property across the dog category, whereas recognition instructions discouraged positive responses to test pictures that differed from previously studied items.

Reasoning and memory judgments could be accounted for by a model in which positive test responses were determined by the total similarity between a test item and previously studied items. Most of the differences between recognition and induction in the rate of "yes" responding were captured by changes in a generalization parameter that increased or decreased the overall psychological distance between study and test stimuli. Although exemplar similarity dominated responses, it was possible to predict responses even better by allowing for subcategory based responses, particularly for induction, as well as deterministic responses, particularly for recognition.

Experiment 1B

Experiment 1A suggested that there is a strong relation between recognition and inductive reasoning. Notably though, exposure to study items in that experiment was relatively brief. It is possible this exposure time was insufficient for people in the induction condition to encode the visual features of each item and generate a category based inference about the target property (i.e., that this property might be shared by all dogs). If short item exposure interfered with the inferential process then the way items were processed in induction would differ little from the encoding process in the recognition condition. Hence, the finding of the close relation between induction and recognition test judgments may be an artifact of the brief stimulus exposure at study. To address this issue, we replicated Experiment 1A with a longer study exposure time.

Method

Eighty undergraduate students from the University of California, Merced campus, participated. Participants were randomly

assigned to one of two conditions: memory ($n = 41$) or inductive reasoning ($n = 39$). The experimental procedure was the same as Experiment 1A, except that pictures were presented for 5 s each during the study phase.

Results and Discussion

The probability of responding "yes" to test items under recognition and induction conditions and associated d' measures of sensitivity are shown in Table 3. Again, we used between-subjects ANOVAs to compare positive responses between tasks, separately for various stimulus types. Overall, the pattern of responding was very similar to Experiment 1A, except that hit rates ("yes" responding to old items) did not differ across task, $F(1, 78) = 1.05$, $p = .31$. Positive responding was again higher for medium dogs than for small dogs, $F(1, 78) = 11.65, p < .001$, partial $\eta^2 = .13$. However, this difference did not interact with task, so small and medium dog responses were again collapsed into a new dogs set. The probability of responding "yes" to new items and lures was higher for induction than recognition, $F(1, 78) = 7.81, p < .001$, partial $\eta^2 = .09$, and $F(1, 78) = 10.19, p < .001$, partial $\eta^2 = .11$, respectively. Sensitivity in the discrimination between old and new items did not differ in recognition and induction, $F(1, 78) = 0.9$, $p = .35$. However, sensitivity in old–lure discrimination was higher for recognition than for induction, $F(1, 78) = 4.39, p < .05$, partial $\eta^2 = .05$.

Again, there was a strong itemwise correlation between the probability of responding positively at test in induction and recognition. The correlation between the recognition and the inductive reasoning conditions was .83 (illustrated by the scatter plot in Figure 2), $p < .001$.

These results replicate most of the key findings of Experiment 1A. Participants doing induction made more positive responses to novel test items and showed poorer discrimination between old and lure items than those doing recognition. Nevertheless, there was a strong positive correlation between the probability of saying "yes" to a test item under induction and recognition conditions. These data confirm that the strong relation between induction and recognition found in Experiment 1 was not just an artifact of the relatively brief presentation of study stimuli.

Modeling

The four GEN-EX models were applied to the 90 data points (45 from each of recognition and induction conditions) in Experiment 1B with the same procedure as in Experiment 1A and the same similarity ratings. The results are summarized in Table 2. Overall, the estimated parameter values for GEN-EX $s + d$ were rather
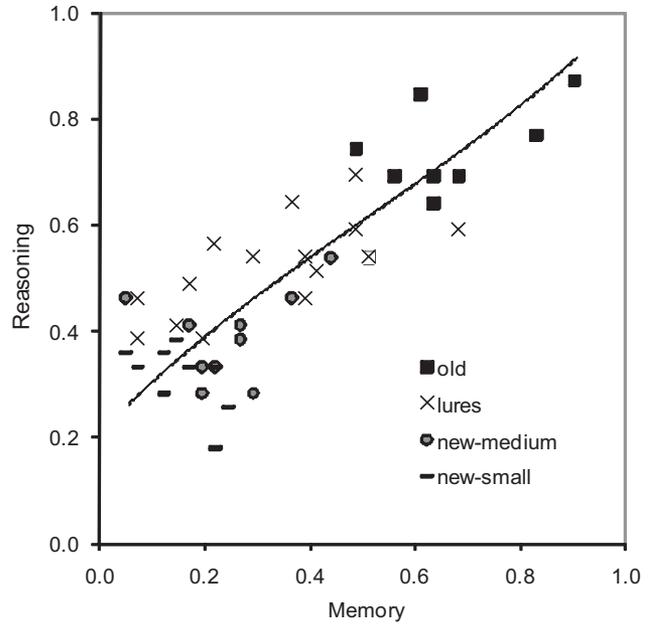

Figure 2. Scatter plots for Experiment 1B, showing memory and reasoning responses across stimuli, with trend line of GEN-EX $s + d$ model predictions. GEN-EX = generalization from examples.

similar to those for Experiment 1A. The β parameter for induction was slightly higher in Experiment 1B than in Experiment 1A, reflecting a somewhat more conservative pattern of responding in Experiment 1B. Once again, the c parameter was higher for recognition than for induction, implying sharper generalization for memory tests. The estimated s parameters suggest a similar level of sub-category based responses as in Experiment 1A. The d parameter for induction was higher for the present experiment than for Experiment 1A, suggesting the longer presentation times facilitated deterministic responding. After taking account of number of free parameters, the GEN-EX $s + d$ model fit significantly better than two of the restricted models, GEN-EX s and GEN-EX null, but the comparison between GEN-EX $s + d$ and GEN-EX d did not quite reach the level of statistical significance.

Key GEN-EX $s + d$ model predictions are shown in Table 3, showing that the model again captured the most important qualitative trends. Likewise, the trend line of model predictions in Figure 2 captures the key trends in the scatter plot of data points. The model predicts a strong correlation between memory and reasoning, .98, even stronger than the correlation in the data, .83.

Table 3
Experiment 1B. Results (Proportion of "Yes" Responses and d') *and Model Predictions*

Result and prediction	Old	New small	New medium	All new	Lure	d' (Old–new)	d' (Old–lure)
Empirical results							
Memory	.67	.16	.25	.20	.33	1.33	0.84
Reasoning	.73	.32	.39	.35	.52	0.98	0.55
Model predictions							
Memory	.67	.14	.23	.18	.34	1.33	0.84
Reasoning	.72	.33	.38	.35	.52	0.97	0.55

Experiment 1C

Experiments 1A and 1B showed a high correlation between memory and reasoning tasks, namely that across the 45 test stimuli, participants' response proportions in a memory test were strongly correlated with another group of participants' response proportions in a reasoning test. Moreover, application of the GEN-EX model showed that the relation was highly systematic. However, it could be argued that a more impressive result would be to show the correlation within the same set of individuals, for then we would be closer to predicting reasoning from memory at the individual level. In addition, we could gain new and potentially informative data, such as what is the conditional probability of responding positively to a reasoning question, given that a participant has or has not responded positively on a memory question. Such results could shed light on the extent to which memory and reasoning performance are stochastically dependent or independent. Brainerd and Reyna (1993) have argued that the relation between memory and reasoning is best evaluated by examining within-subjects dependencies in performance on both tasks. Such an approach has revealed that reasoning in tasks like class inclusion and transitive inference can operate independently of memory for the precise details of premises in an argument.

In Experiment 1C, therefore, we asked each participant to answer both a memory question and a reasoning question for each test item. To minimize confusion and to allow for a systematic analysis of possible carryover effects, half the participants were asked a memory question and then a reasoning question after the presentation of each test item, and half the participants were asked a reasoning question and then a memory question after the presentation of each test item.

Method

The method, including the materials and the procedure for the study phase, was the same as Experiment 1A, except for the following. Seventy-nine University of California, Merced, students participated: 39 in the memory-first condition and 40 in the reasoning-first condition.

At the beginning of the study phase, participants were given neutral instructions that did not indicate the nature of the test questions: "In this experiment, you will see a set of pictures of animals. Please pay close attention to these pictures. In the second part of the experiment, you will be asked questions about the pictures." At the beginning of the test phase, participants were informed that each animal they had seen was discovered to have something called beta cells. Then they were informed that they would be asked two questions per test item: whether they had seen the animal and whether they thought it had beta cells. In the test phase, participants were asked two questions in sequence, for each of the 45 test items. In the memory-first condition, the memory question was always asked first, and in the reasoning-first condition, the reasoning question was asked first.

Results

A preliminary analysis showed that the order of question presentation had no significant effect on "yes" responding or d' sensitivity ($Fs < 2.2$), so the data were collapsed across this variable. The resulting probabilities of responding "yes" to test items under recognition and induction conditions and associated d' measures of sensitivity are given in Table 4, with "yes" responding to small and medium dogs collapsed into a new dogs set. The table shows that there was a much higher rate of positive responding to all classes of test items in the induction condition than in the recognition condition; old, $F(1, 77) = 9.67$, $p = .003$, partial $\eta^2 = .11$; new, $F(1, 77) = 29.04$, $p < .001$, partial $\eta^2 = .27$; lures, $F(1, 77) = 72.21$, $p < .001$, partial $\eta^2 = .48$. Sensitivity to both old–new, $F(1, 77) = 61.53$, $p = .003$, partial $\eta^2 = .32$, and old–lure discriminations, $F(1, 77) = 65.32$, $p = .003$, partial $\eta^2 = .44$, was higher for recognition than for induction.

Critically, the itemwise correlation between "yes" responding in induction and recognition questions remained high and positive $r(43) = 0.84$, $p < .001$. Following the approach of Brainerd and Reyna (1993), we also calculated the conditional probability of making a positive response to a test item in induction, given that the item had been recognized, p("yes" induction|"yes" recognition) = .95. This was substantially higher than the conditional probability of a yes response in induction given a "no" response in recognition, p("yes" induction|"no" recognition) = .31. These results support the assumption of stochastic dependence between induction and recognition test judgments.

Modeling

Due to the lack of notable differences between the memory-first and reasoning-first conditions, we pooled these together, yielding 90 data points to be modeled for Experiment 1C, 45 from memory judgments and 45 from reasoning judgments.

The four versions of the GEN-EX model were applied to these data. Overall, the estimated parameter values for GEN-EX $s + d$ were similar to those for Experiments 1A and 1B. Perhaps the key difference was that the higher level of discrimination in the memory condition was accomplished with a high value of the deterministic responding parameter, d. With the high value of d for

Table 4
Experiment 1C. Results (Proportion of "Yes" Responses and d'*) and Model Predictions*

Result and prediction	Old	New small	New medium	All new	Lure	d' (Old–new)	d' (Old–lure)
Empirical results							
Memory	.81	.09	.19	.14	.28	1.94	1.44
Reasoning	.85	.17	.36	.26	.64	1.68	0.67
Model predictions							
Memory	.80	.08	.17	.13	.29	1.99	1.41
Reasoning	.88	.18	.34	.26	.62	1.82	0.88

memory, it was not necessary to also have a higher value of c for the memory condition. Note that when the GEN-EX s model was applied to the data, without a deterministic responding component, c was higher for memory, 3.64, than for reasoning, 3.00. There was some evidence for subcategory-based reasoning, with an s value of .08 for the reasoning condition. After taking account of number of free parameters, the GEN-EX $s + d$ model fit significantly better than two of the restricted models, GEN-EX s and GEN-EX null, but the comparison between GEN-EX $s + d$ and GEN-EX d did not quite reach the level of statistical significance.

Key GEN-EX $s + d$ model predictions are shown in Table 4, showing that the model again captured the most important qualitative trends. Likewise, the trend line of model predictions in Figure 3 captures the key trends in the scatter plot of data points. In this figure, it appears that the data points for old items stand apart from the data points for lure, new–medium, and new–small items. The model was able to capture this result with the high d value for the memory condition. Finally, the model predicts a strong correlation between memory and reasoning, .91, even stronger than the correlation in the data, .84.

Discussion

These data replicate and extend the results of the first two experiments. When both judgments were made by the same participants, there was a higher rate of positive responding to induction than to recognition questions for all types of test items. There were also robust differences between induction and recognition in sensitivity to old–new and old–lure discriminations. If anything, the results were stronger when induction and recognition judgments were manipulated within-subjects. Notably, the within subjects data also supported previous findings of a strong positive relation between recog-

Figure 3. Scatter plots for Experiment 1C, showing memory and reasoning responses across stimuli, with trend line of GEN-EX $s + d$ model predictions. GEN-EX = generalization from examples.

nition and induction. The probability of judging that a test item had beta cells increased dramatically when the item was recognized. For comparison, note that Metcalfe and Fisher (1986) found evidence for independence between recognition and categorization.

These results support the conclusion that a consideration of exemplar similarity underpins both induction and recognition judgments, although note that exemplar models may be flexible enough to account for a variety of patterns of contingency between tasks (Nosofsky, 1988a). This conclusion was further supported by the modeling results that show that an exemplar model gives a good account of both induction and recognition data. As in the previous studies, however, model fit was improved by adding supplementary rule-based parameters to GEN-EX (sub-typing in induction; identity-based responding in recognition).

Experiment 1D

The previous experiments showed that when induction and recognition tasks are made comparable, there is high degree of correspondence between the test phase judgments in each task. We have argued that this reflects the central role of exemplar similarity in both tasks. There is, however, an alternative explanation. In Experiments 1A, 1B, and 1C, both recognition and induction had a significant memory component. In both tasks, participants had to study a set of stimuli, retain these over 60 s, and then compare their memory representations with test phase stimuli. By doing this, we may have inadvertently converted the induction task into a type of recognition memory, setting the stage for high correlations between reasoning and memory.

To address this issue, we ran an additional induction condition that had no memory component. In this case, study phase stimuli were available for inspection during test. The data from this new condition were compared with those from the recognition condition in Experiment 1A. If the close relation between memory and induction found in the previous studies was an artifact of a common memory component then this relation should be substantially weakened when the new induction condition is correlated with recognition performance. If, however, it is the assessment of exemplar similarity that underlies this relation then it should be preserved when induction does not involve memory.

Method

Forty University of California, Merced, students participated. The procedure for the no memory induction condition was like the inductive reasoning condition in Experiment 1A, except that the study and test phases were conducted simultaneously. The 10 study items were presented in two rows of five tiles in the top half of the computer screen. The positions of study items were randomized for each participant, but the positions stayed the same as the experiment was conducted on each individual. It was indicated that all of the study items had beta cells. On each of the 45 test trials, one test item was presented in the bottom half of the screen as the participant was asked whether that animal has beta cells. The pictures were the same size as in the previous experiments.

Results

The probability of responding "yes" to test items in the no-memory induction condition and associated d' measures of sensi-
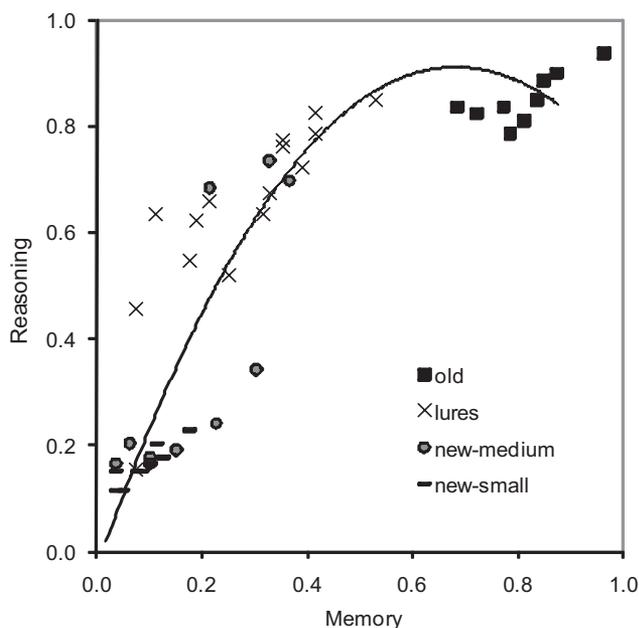
tivity are given in Table 5. These data were compared with the induction and recognition memory conditions in Experiment 1A in separate analyses.

Removing the memory component from induction altered the probability of responding "yes" during the test phase. Compared with the induction condition in Experiment 1A, those in the no memory induction condition were more likely to respond positively to old items, $F(1, 70) = 8.31$, $p < .01$, partial $\eta^2 = .11$, and less likely to respond positively to new medium dogs and lures, $F(1, 70) = 4.2$, $p < .05$, partial $\eta^2 = .06$, and $F(1, 70) = 16.01$, $p < .001$, partial $\eta^2 = .19$, respectively. This led to significantly higher levels of sensitivity in discrimination between old and new items, $F(1, 70) = 8.55$, $p < .01$, partial $\eta^2 = .11$, and between old and lure items, $F(1, 70) = 27.19$, $p < .001$, partial $\eta^2 = .28$, in the no memory condition, as compared with the original induction group.

The probability of responding positively to old items in the no memory induction condition was also significantly higher than in the recognition condition of Experiment 1A, $F(1, 70) = 44.35$, $p < .001$, partial $\eta^2 = .39$. In the case of medium dogs, though, the rate of positive responding was still reliably higher for no memory induction than for recognition, $F(1, 70) = 4.06$, $p < .05$, partial $\eta^2 = .06$. There were no significant differences between the induction–no memory and recognition conditions for sensitivity in old–new discrimination. Sensitivity in old–lure discrimination was significantly higher for no memory induction than for recognition, $F(1, 70) = 8.06$, $p < .01$, partial $\eta^2 = .11$.

Despite these changes in the probability of positive responding, there was still a strong itemwise correlation between test phase responding in the no memory induction condition and recognition, $r(43) = .86$. We plot this relation in Figure 4, with induction responses from Experiment 1D on the *y* axis and recognition responses from Experiment 1A on the *x* axis. What is most distinctive about this scatter plot is the separation between old items and the remaining items, as if there was some tendency to respond deterministically and positively on old items.

Modeling

The four versions of the GEN-EX model were applied to the 45 data points in Experiment 1D, as summarized in Table 2. The GEN-EX $s + d$ model fit significantly better than each of the three restricted models. What is most notable about the parameter estimates is the high level of d for induction, indicating a high level of deterministic responding. Because the study items were available at the time of test, there was a strong tendency to respond deterministically based on item matches.

Key model predictions are shown in Table 5 and Figure 4, showing that the GEN-EX $s + d$ model captured the important
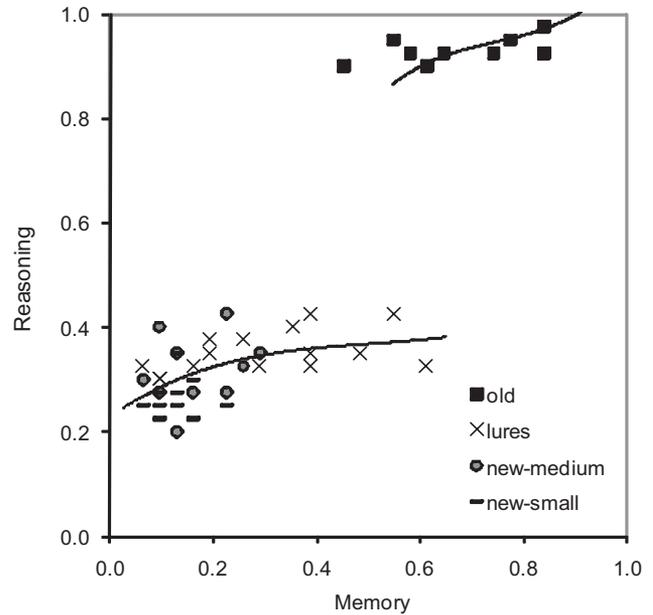

Figure 4. Scatter plots for Experiment 1D, showing memory (from Experiment 1A) and reasoning responses across stimuli, with trend line of GEN-EX $s + d$ model predictions. GEN-EX = generalization from examples.

qualitative trends (note that two trend lines are plotted in Figure 4, due to the bimodal pattern of model predictions as a consequence of the very high level of the d parameter). The model predicted a .94 correlation between induction responses in Experiment 1D and recognition responses in Experiment 1A.

Discussion

When the memory component was removed from induction, the most notable changes in test phase responding were that positive responses to old items were close to ceiling, and positive responses to novel items declined relative to the induction and recognition conditions in Experiment 1A. In many respects, these findings are unsurprising. When old test items were presented on the same screen with identical study items that were known to have beta cells, people generalized the property to these test items. Notably though, there was still a tendency for those performing induction to make more positive responses to some novel items (medium dogs) than those doing recognition, even when they did not have to rely on memory for study instances. Most notably, the relation between positive responding at test in induction and recognition

Table 5
Experiment 1D. Results (Proportion of "Yes" Responses and d′*) and Model Predictions*

Result and prediction	Old	New small	New medium	All new	Lure	d′ (Old–new)	d′ (Old–lure)
Empirical results							
Reasoning	.93	.27	.32	.29	.35	2.03	1.86
Model predictions							
Reasoning	.93	.28	.31	.29	.35	2.02	1.85

was as strong (if not stronger) when memory was removed from induction. This result suggests that the close relation between recognition and induction judgments found in the past three studies was not an artifact of the presence of a memory component in both tasks.

The GEN-EX $s + d$ model was able to capture the main trends in induction responding when there was no need to retrieve exemplars from memory. As in the previous experiments, the model still predicted a broader generalization gradient for induction than recognition. The model also successfully captured the main change in induction when the memory component was removed, namely, the marked increase in deterministic induction based on matches with old items. This reinforces the value of adding the deterministic d parameter to the GEN-EX framework; this clearly allows the model to capture inductive responding across a wider range of presentation contexts.

Experiment 2A

Experiments 1A–1D showed that when comparable reasoning and memory tasks are used, there is a close relation between how people make recognition and induction judgments. The application of GEN-EX suggested that data from both tasks could be accommodated within a single exemplar model. The aim of the remaining experiments was to examine the generality of the recognition–induction relation and an exemplar model of reasoning and memory. In each experiment, we varied task factors that had the potential to differentially affect recognition or induction and examined whether the empirical relation between judgments on these two tasks varied as a function of these manipulations. We also examined whether the fit of GEN-EX differed across task conditions.

In Experiment 2A, we examined the effect of presentation frequency on memory and reasoning. In effect, this was a strength manipulation that we hoped to be more effective than the presentation duration manipulation in Experiment 1B. The frequency manipulation was that each study item was presented one time (as in Experiments 1A–1D) or three times. Manipulation of presentation frequency has known effects on memory. Increasing the frequency of presentation of individual study items increases the sensitivity of old–new discrimination in recognition (Flexser & Bower, 1974; Glanzer, Kim, Hilford, & Adams, 1999; Hintzman, 2001). In terms of the GEN-EX model, such effects could be explained by increases in study item presentation frequency leading to more elaborate processing of the distinctive features of study stimuli with a resulting decrease in the perceived similarity of novel test items. Increases in study item frequency are also likely to narrow the generalization parameter around study items.

The effects of varying presentation frequency on induction are more difficult to predict and depend on how the induction process is conceptualized. If we are correct in assuming that induction is affected by changes in old–new similarity in much the same way as recognition then we should also see a sharper gradient of inductive generalization (i.e., higher ds) as the presentation frequency of study items increases. Tenenbaum and Griffiths (2001) made a related prediction based on the size principle, namely, the idea that more specific hypotheses are favored over more general hypotheses. This prediction was confirmed in a word learning task in which it was found that when a novel label was applied several

times to a given object, children and adults showed narrower generalization in their extension of the label to similar objects than when the label was only presented once (Xu & Tenenbaum, 2007). One consequence of the size principle is that confidence in a narrow range of generalization will be increased even further when there are more observations within a narrow range.

Alternately, there is some suggestion from previous work that increasing the presentation frequency of items that share a given property might have the opposite effect on induction. The monotonicity effect is a well documented phenomenon whereby as the number of instances known to share a property increases, there is an increase in the probability of the property being generalized to other category members (e.g., Feeney, 2007; Osherson et al., 1990; Rotello & Heit, 2009). The current study phase manipulation differs from previous studies of monotonicity in that items with the same property were repeated rather than new items being added. Nevertheless, repetition could increase the opportunity for the abstraction of common characteristics across the dog study items and promote the generalization of the target property to test items that share many features in common with these study items.

Method

Participants. One hundred and twenty-one University of California, Merced, students participated, with 31 randomly allocated to the Memory-Frequency 1, and 30 each to the Reasoning-Frequency 1, Memory-Frequency 3, and Reasoning-Frequency 3 conditions.

Procedure. For the Frequency-1 conditions, the method was exactly the same as Experiment 1A. The only change for the Frequency 3 conditions was to present each studied stimulus three times, randomly interspersed through the list, rather than once.

Results

The probability of responding "yes" to test items under recognition and induction conditions is shown in Table 6. We used ANOVAs to compare positive responses, with task and frequency as between-subject variables, separately for various stimulus types. Unlike previous studies, positive responses to old items tended to be higher for recognition than for induction, but this difference did not reach significance, $F(1, 116) = 3.73$, $p = .06$. As in previous studies, there were more "yes" responses to new medium than small dogs, $F(1, 116) = 43.9$, $p < .001$, partial $\eta^2 = .2$, but this effect did not interact with task condition or study phase frequency. Hence, as before, responses to small and medium test dogs were collapsed. The probability of responding "yes" to these new dogs and to lures was higher for induction than recognition, $F(1, 116) = 18.13$, $p < .001$, partial $\eta^2 = .14$; and $F(1, 116) = 7.97$, $p < .01$, partial $\eta^2 = .06$, respectively). Frequency of presentation of study phase items had no significant effect on positive responses to old, new, or lure items and did not interact with task ($Fs < 1.5$).

Individual ds were calculated (see Table 6) and analyzed as in previous studies. Sensitivity in the discrimination between old and new items was higher overall for recognition than induction, $F(1, 116) = 18.26$, $p < .001$, partial $\eta^2 = .14$. In both recognition and induction conditions, sensitivity in the old–new discrimination tended to be higher in Frequency 3, but this difference did not reach significance, $F(1, 116) = 2.83$, $p = .09$, partial $\eta^2 = .03$.

Table 6
Experiment 2A. Results (Proportion of "Yes" Responses and d′*) and Model Predictions*

Result and prediction	Old	New small	New medium	All new	Lure	d′ (Old–new)	d′ (Old–lure)
Results							
Frequency 1 memory	.79	.18	.34	.26	.46	1.47	0.92
Frequency 1 reasoning	.73	.39	.51	.45	.52	0.73	0.56
Frequency 3 memory	.84	.18	.31	.34	.36	1.67	1.33
Frequency 3 reasoning	.76	.34	.42	.38	.52	1.01	0.68
Predictions							
Frequency 1 memory	.79	.17	.31	.24	.47	1.51	0.88
Frequency 1 reasoning	.73	.40	.48	.44	.53	0.76	0.52
Frequency 3 memory	.83	.18	.31	.20	.40	1.82	1.21
Frequency 3 reasoning	.76	.32	.43	.37	.52	1.04	0.65

There was no interaction between task condition and study frequency, $F(1, 116) = 0.34$. A similar pattern was found for sensitivity in the old–lure discrimination, in which sensitivity was higher for recognition than induction, $F(1, 116) = 15.63, p < .001$, partial $\eta^2 = .12$, and higher for Frequency 3 than Frequency 1, $F(1, 116) = 4.42, p < .05$, partial $\eta^2 = .04$, with no interaction between task and frequency, $F(1, 116) = 0.51$.

The proportion of positive responses for each of the 45 test items was again averaged across participants, and correlations between responses in the recognition and induction conditions were computed separately for the Frequency 1 and Frequency 3 groups. As in previous studies, test responses in recognition and induction were strongly correlated; Frequency 1: $r(43) = .82$; Frequency 3: $r(43) = .84$, $ps < .001$. This relation is illustrated by the scatter plots in Figure 5.

Modeling

The four versions of the GEN-EX model were applied to the 180 data points in Experiment 2A with the same procedure as in the previous experiments and the same similarity ratings. Different parameters were used for the Frequency 1 and Frequency 3 conditions, except for the c parameter for similarity ratings. Although it would have been possible to model the Frequency 3 conditions by representing each studied item three times in memory (e.g., Heit, 1994, 1998b; Nosofsky, 1988b), the consequence would simply be to make all of the familiarity scores derived from Equation 1 three times as large. In Equation 2, the β parameters would then be estimated to be three times as large, with the result that the model would make exactly the same predictions.

The results of the modeling are summarized in Table 2. The average estimated parameter value of c for similarity ratings was 1.80. For the GEN-EX $s + d$ model, the c parameters were higher for memory than for reasoning, implying a sharper generalization gradient. For both recognition and induction, increasing presentation frequency led to an increase in the c parameter, indicating that more presentations sharpened generalization for both memory and reasoning. The best fitting model actually estimated zero influence of subcategories, hence the GEN-EX $s + d$ and GEN-EX d predictions were equivalent. However, the fit of GEN-EX $s + d$ was significantly better than that of GEN-EX s and GEN-EX null, indicating that including

a component for deterministic responding significantly improved performance of the model.

Key model predictions are shown in Table 6, showing that the model again captured the most important qualitative trends such as a higher false alarm rate for reasoning than memory, higher sensitivity for Frequency 3 than Frequency 1, and the relative positions of old, lure, new medium, and new small stimuli. Likewise, trend lines for model predictions captured the key trends in the scatter plots of data points (see Figure 5). The model predicts a strong correlation between memory and reasoning, .98, across 90 data points for memory and 90 data points for reasoning, even stronger than the correlation in the data, .83.

Discussion

This study replicated many of the key results of earlier studies. Although sensitivity was higher for recognition than induction, there was a robust positive relation between the probability of saying "yes" to test items in each task. As in many previous studies of recognition (e.g., Hintzman, 2001), increasing the frequency of presentation of study phase items led to an increase in sensitivity in old–new discrimination as measured by $d′$. The interesting and relatively novel finding was that presentation frequency affected inductive generalization in the same way. When study phase dogs with a novel anatomical property were repeated three times, people were less likely to generalize the property to other dogs than when study exemplars were only presented once.

The parallel effect of frequency on recognition and induction is consistent with the view that exemplar similarity is a common core process in both of these tasks and that the frequency manipulation affected the way that similarity was computed, as reflected by the changes in the c parameters. This conclusion is also supported by the modeling results in which the main trends in induction and recognition, under one and three presentations of study items, were well captured by the GEN-EX model. Both the differences between induction and recognition and between different levels of study frequency could be accounted for by changes in the c parameter of GEN-EX.

These findings suggest that the repetition of category members that share a property has a very different effect on induction than increasing the number of discrete instances that share the property (i.e., premise monotonicity). Previous studies of monotonicity
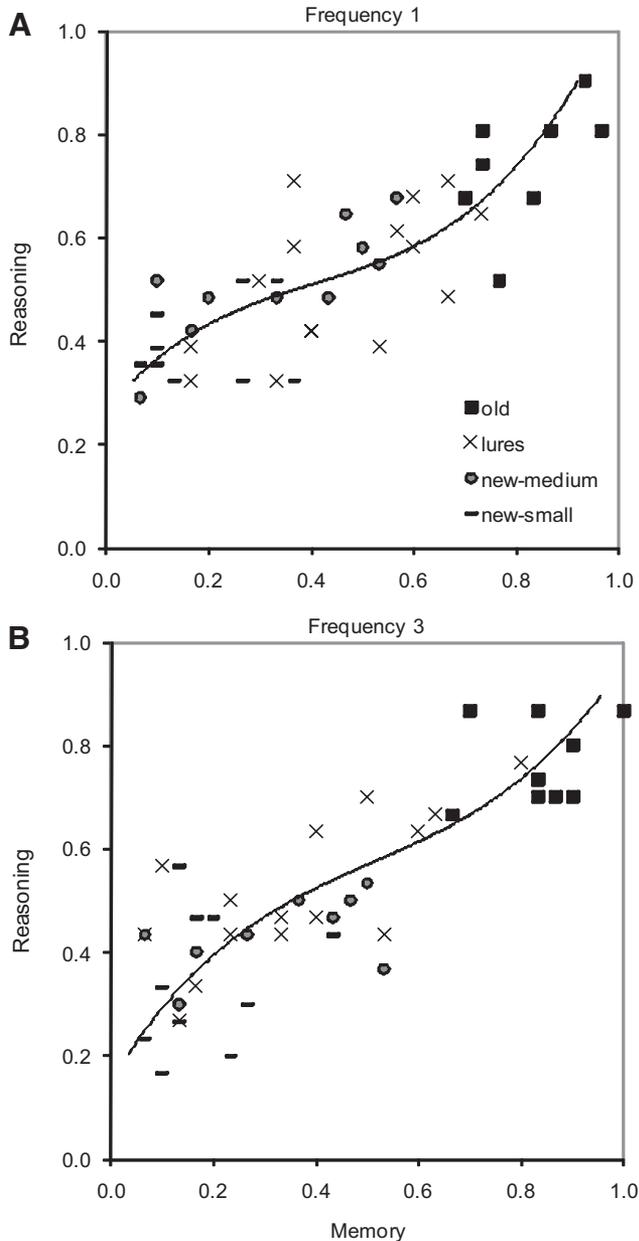
Figure 5. Scatter plots for Experiment 2A, showing memory and reasoning responses across stimuli, for (A) Frequency 1 and (B) Frequency 3, with trend lines of GEN-EX $s + d$ model predictions. GEN-EX = generalization from examples.

suggest that increasing the number of category members that share a property facilitates extraction of the common features of those instances and promotes property generalization to novel test items that share those features (Sloman, 1993). Our data, however, suggest that item repetition at study enhances the encoding of distinctive features that distinguish study and test items (cf. Nosofsky, 1988b).

We note that there was no evidence for subcategory-based responding in Experiment 2A. We did not have a strong prediction that this pattern of response would always be present; we have

simply included this possibility in the modeling, as a means of assessing when it is there and, potentially, whether it varies between induction and recognition tasks.

Experiment 2B

Experiment 2B was an attempt to have a somewhat stronger manipulation of frequency than in Experiment 2A. For example, in Experiment 2A, presenting stimuli three times rather than once led to only a 5% increase in the hit rate for the memory condition and a 3% increase for the reasoning condition. Here, we manipulated frequency within list rather than between list. Previous work has shown that within-list repetition of items results in a significant increase in the retrieval accuracy for those items (e.g., Criss, 2009; Ratcliff, Clark, & Shiffrin, 1990). Hence, within the study list, half the pictures were presented once and half were presented three times. We were interested in whether this potentially stronger manipulation of frequency would have the same or different effects on memory and reasoning and whether the GEN-EX model could accommodate all of the results.

Method

Participants.   One hundred and sixty-eight students from the University of New South Wales participated for course credit, with equal numbers randomly allocated to the memory and reasoning task conditions.

Procedure.   The method was generally the same as Experiment 1A, except that half of the study items were presented three times, whereas the remainder were presented only once. Half the participants were assigned to study Version A, in which five out of the 10 study items (chosen arbitrarily) were presented three times. In study Version B, the other 5 study items were presented three times. The order of presentation of all study items was randomized for each participant. The test procedure was identical to Experiment 1A.

Results and Discussion

Preliminary analyses did not find any differences in hit or false alarm rates for different item types between participants presented with the different study sets, so data were pooled across this factor for the next set of descriptive analyses. As in previous studies, there were more positive responses to new medium dogs than small dogs, $F(1, 164) = 50.84$, $p < .001$, partial $\eta^2 = .24$, but this effect did not interact with task condition or study phase frequency. As before, responses to small and medium test dogs were pooled.

The probability of responding "yes" to test items under recognition and induction conditions is shown in Table 7. Hit rate data were entered into a 2 (task) $\times$ 2 (Frequency 1 vs. Frequency 3) analysis with repeated measures on the second variable. Hit rates were higher for old items that had been presented three times during the study phase ($M = .91$) than for those presented once ($M = .81$), $F(1, 164) = 45.56$, $p < .001$, partial $\eta^2 = .22$. This effect was qualified by an interaction with task condition, $F(1, 164) = 5.06$, $p = .03$, partial $\eta^2 = .03$. Table 7 shows that the effect of frequency on hit rates was larger in the recognition than in the induction condition. A one-way ANOVA examining the

Table 7
Experiment 2B. Results (Proportion of "Yes" Responses and d'*) Based on Study Items Presented One Time, Study Items Presented Three Times, and Model Predictions*

Result and prediction	Old (×1)	Old (×3)	New small	New medium	All new	Lure	d' (Old × 1–new)	d' (Old × 3–new)	d' (Old × 1–lure)	d' (Old × 3–lure)
Empirical results										
Memory	.79	.93	.13	.22	.18	.32	1.84	2.18	1.34	1.68
Reasoning	.82	.89	.27	.39	.34	.50	1.40	1.56	0.82	0.98
Predictions										
Memory	.80	.90	.08	.20	.14	.36	1.89	2.32	1.17	1.60
Reasoning	.81	.88	.26	.39	.32	.53	1.33	1.66	0.80	1.13

effects of task on positive responses for new items found a higher false alarm rate in the induction than in the recognition condition, $F(1, 164) = 19.85$, $p < .001$, partial $\eta^2 = .11$. False alarm rate for lures were also higher for induction than recognition, $F(1, 164) = 25.0$, $p < .001$, partial $\eta^2 = .13$.

Separate individual *d*s were calculated with hit rates for study items presented once and for those presented three times (see Table 7) and analyzed in a 2 (task) × 2 (Frequency 1 vs. Frequency 3) analysis with repeated measures on the second factor. Sensitivity in the discrimination between old and new items was higher when old items were repeated three times during study ($M_{OLD \times 1} = 1.62$ vs. $M_{OLD \times 3} = 1.87$), $F(1, 164) = 40.93$, $p < .001$, partial $\eta^2 = .20$, and higher in the recognition ($M = 2.01$) than in the induction condition ($M = 1.5$), $F(1, 164) = 14.78$, $p < .001$, partial $\eta^2 = .08$. There was also a significant interaction between these factors, $F(1, 164) = 4.86$, $p = .03$, partial $\eta^2 = .03$. Table 7 shows that the manipulation of study frequency had a larger effect on recognition than induction. A similar pattern was found for sensitivity in the old–lure discrimination with sensitivity higher for repeated study items ($M_{OLD \times 1} = 1.08$ vs. $M_{OLD \times 3} = 1.33$), $F(1, 164) = 40.93$, $p < .001$, partial $\eta^2 = .17$, and higher for recognition ($M = 1.51$) than induction ($M = 0.9$), $F(1, 164) = 21.56$, $p < .001$, partial $\eta^2 = .17$. Again, these factors interacted such that repeating study items had a greater effect on sensitivity in recognition than in induction, $F(1, 164) = 4.86$, $p = .03$, partial $\eta^2 = .03$.

As in previous studies, the itemwise correlation between test responses in recognition and induction was high and positive, for Version A, $r(43) = 0.93$, $p < .001$, and for Version B, $r(43) = 0.92$, $p < .001$. See Figure 6 for scatter plots of the results.

Modeling

The four versions of the GEN-EX model were applied to the 180 data points in Experiment 2B, here considering the two study lists separately. The parameters to be estimated were the same as in Experiment 2A, with the key difference being that studied items were either represented one time or three times in memory for Versions A and B. The results are summarized in Table 2. For the GEN-EX $s + d$ model, the estimated parameters are rather similar to those of Experiment 2A. (Note that it was not possible to explain the frequency manipulation just by varying the c parameters; representing some stimuli multiple times was crucial.) The best fitting model estimated a near-zero influence of subcategories, hence the GEN-EX $s + d$ and GEN-EX d predictions were almost equivalent. However, the fit of GEN-EX $s + d$ was significantly

better than that of GEN-EX s and GEN-EX null, indicating that including a component for deterministic responding significantly improved performance of the model.

Key model predictions are shown in Table 7, showing that the model again captured the most important qualitative trends, such as a substantially higher hit rate for old three items than for old one items, a higher false alarm rate for reasoning than memory, and the relative positions of old, lure, new medium, and new small stimuli. Likewise, trend lines for model predictions captured the key trends in the scatter plots of data points (see Figure 6). The model predicts a strong correlation between memory and reasoning, .98, across 90 data points for memory and 90 data points for reasoning, even stronger than the correlation in the data, .94, across these same 90 data points.

Taken together, Experiments 2A and 2B suggest that frequency manipulations have similar effects on memory and reasoning, in both cases increase the probability of saying "yes" to old items as well as ability to distinguish old items from new items. The GEN-EX model can account for the effects of this manipulation without additional assumptions.

Experiment 3

In this experiment, we examined the effects of adding items from outside the target category during the study phase on subsequent induction and recognition judgments. As noted previously, a robust finding in the induction literature is that increasing the number of categories (or category members) that share some property leads to wider generalization of that property, a phenomenon referred to as "premise monotonicity" (Feeney, 2007; Osherson et al., 1990; Sloman, 1993). Hence, adding study items from a variety of animal categories that all share a target property should lead participants to generalize that property more broadly. If birds and fish as well as dogs have beta cells then this might seem reasonable grounds to generalize the property to other animals. This would mean a very high rate of "yes" responding to novel dogs at test. By comparison it is not clear that adding study items that are perceptually discriminable from the target set of dogs should radically alter recognition responses to new dogs at test.

Again, however, the GEN-EX model suggests a different set of predictions. According to GEN-EX, study exposure to dogs and items from other categories should increase the relative similarity of test dogs to study dogs (cf. diagnosticity effects in Tversky, 1977). This should lead to more "yes" responses (and poorer test phase sensitivity) in both induction and recognition conditions.
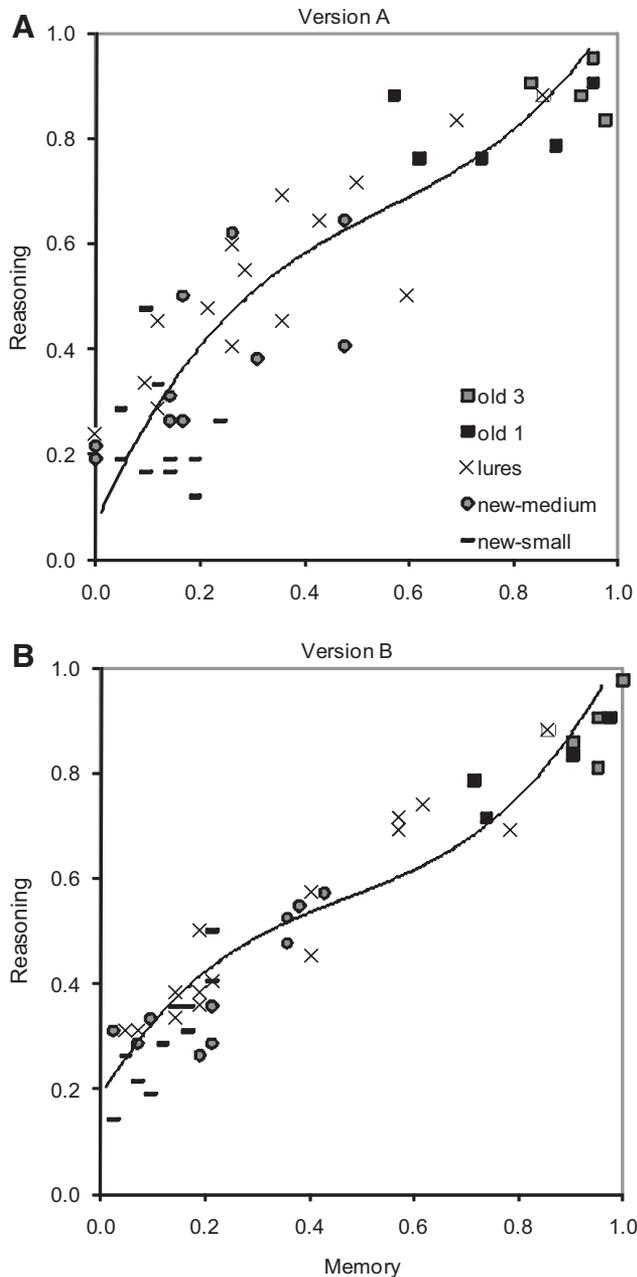
Figure 6. Scatter plots for Experiment 2B, showing memory and reasoning responses across stimuli, for (A) Version A and (B) Version B, with trend lines of GEN-EX *s* + *d* model predictions. GEN-EX = generalization from examples.

Method

Participants. A total of 176 undergraduates from the University of California, Merced, participated. Forty-four participants were randomly allocated to the four conditions generated by the factorial crossing of task (induction vs. recognition) and study list (dogs only vs. dogs + nondogs).

Procedure. The study phase procedure for the recognition and induction conditions was similar to previous experiments

except that in this case, 15 dog photographs were shown during study. These were the 15 large dogs previously used as lure test items. The reason for increasing the number of dog study items was that we increased the total number of study items, adding 15 birds and 15 fish in the dogs + nondogs conditions, and we wanted the key dog stimuli to still be a substantial part of the study list. In the dogs + nondogs conditions, 45 pictures (15 dogs, 15 birds, 15 fish) were presented for 2 s each during the study phase in random order. In the dogs-only conditions, only the 15 dog pictures were presented. At test, all groups made judgments about 45 dogs (15 study, 10 new small, 10 new medium, and 10 lures). The lure items were the 10 pictures of dogs used as old items in previous studies. Induction and recognition instructions at study and test were manipulated as in previous studies.

Results

The probability of responding positively at test is shown in Table 8. As in previous studies, there were more positive responses to new medium dogs than small dogs, $F(1, 172) = 74.67$, $p <$ .001, partial $\eta^2 = .30$. However this difference did not interact with either of the group variables ($Fs < 2.0$), so in subsequent analyses, responding was collapsed across these two sets. Positive responding for old, new, and lure items was analyzed with factorial analyses of variance with task (recognition vs. induction) and study list (dogs vs. dogs + nondogs) as between-subjects factors. There were no main effects of task or study list on "yes" responses to old items, but there was a significant interaction between these variables, $F(1, 172) = 4.11$, $p < .05$, partial $\eta^2 = .02$. Simple effects tests showed that "yes" responding to old items was reliably higher for recognition than induction in the dogs only condition, $t(86) = 2.04$, $p < .05$, but there was no difference between recognition and induction in the dogs + nondogs condition, $t(86) = -0.57$, $p = .57$. The probability of responding "yes" to new items and lures was higher for induction than recognition, $F(1, 172) = 8.24$, $p < .01$, partial $\eta^2 = 0.05$; $F(1, 172) = 18.32$, $p < .001$, partial $\eta^2 = 0.10$, respectively. There were no main effects of study list for these items but the study list factor did interact with task for responses to new items, $F(1, 172) = 4.3$, $p < .05$, partial $\eta^2 = 0.02$. Table 8 shows that the difference in "yes" responding between the recognition and induction conditions was larger following the dogs + nondogs study list than the dogs only list.

Individual *d*s were calculated for discrimination between old and new items and between old and lure items. These were analyzed with factorial analyses of variance with task and study list as between-subjects factors. Sensitivity in the discrimination between old and new items was higher for recognition than induction, $F(1, 172) = 10.99$, $p < .001$, partial $\eta^2 = .06$, and higher following exposure to the dogs only, as compared with the dogs + nondogs set, $F(1, 172) = 7.63$, $p < .05$, partial $\eta^2 = .04$. Old–lure discrimination sensitivity was higher for recognition than induction, $F(1, 164) = 8.3$, $p < .01$, partial $\eta^2 = .12$. No other significant main effects or interactions were found ($Fs < 1.0$).

Item-wise correlations between positive responses in induction and recognition tests were high and positive for both the dogs only condition, $r(43) = 0.83$, $p < .001$, and the dogs + nondogs condition, $r(43) = 0.91$, $p < .001$, (see Figure 7 for scatter plots).

Table 8
Experiment 3. Results (Proportion of "Yes" Responses and d′*) and Model Predictions*

Result and prediction	Old	New small	New medium	All new	Lure	d' (Old–new)	d' (Old–lure)
Results							
Dogs only memory	.77	.26	.38	.32	.38	1.23	1.05
Dogs only reasoning	.67	.29	.40	.35	.53	0.82	0.36
Dogs + nondogs memory	.65	.23	.37	.29	.30	0.94	0.92
Dogs + nondogs reasoning	.69	.43	.50	.46	.50	0.60	0.50
Predictions							
Dogs only memory	.77	.26	.36	.31	.31	1.23	1.05
Dogs only reasoning	.68	.33	.42	.38	.47	0.82	0.36
Dogs + nondogs memory	.65	.23	.32	.27	.35	0.98	0.76
Dogs + nondogs reasoning	.68	.41	.49	.45	.53	0.60	0.38

Modeling

This experiment had a different and larger study set than the previous experiments (i.e., 15 dogs rather than 10 dogs). Modeling the responses with the same method as in the previous experiments would require pairwise similarity ratings for 675 combinations (15 study items × 45 test items) for just the dog stimuli and 2025 combinations (45 study items × 45 test items) if the bird and fish stimuli were included as well. Instead, we took the approach of Heit and Hayes (2008) and simulated the overall patterns of responses, based on randomized configurations of hypothetical stimuli. Because the model was fitted to average response rates for different item types rather than to individual item responses, there were fewer independent data points and less opportunity to add free parameters to the GEN-EX model. Hence, only the GEN-EX null model was applied. Given that there was little obvious perceptual similarity between birds and fish at study and dogs at test, we did not directly include the bird and fish study items in the simulations (although their impact could be observed in parameter changes).

We made predictions for four types of stimuli, assigned by simulations to randomized positions in hypothetical two-dimensional stimulus space. The study list contained 15 large dogs. Their *XY* positions were drawn randomly from a bivariate normal distribution with mean of (0, 0) and a standard deviation of 1. The test list included these 15 old items as well as 10 lure items, additional large dogs also drawn randomly from a bivariate normal distribution with a mean of (0, 0) and a standard deviation of 1. There were 10 new items that were medium dogs, in the simulations drawn randomly from a bivariate normal distribution with an origin of (A, A) and a standard deviation of 1. Finally, there were 10 new items that were small dogs in the simulations drawn randomly from a bivariate normal distribution with an origin of (B, B) and a standard deviation of 1. Note that A and B were estimated as free parameters—these values would reflect the average positions in psychological space of the large, medium, and small dogs. It was expected that B > A, reflecting the notion that large dogs are more similar to medium dogs than large dogs are to small dogs.

The model was used to make predictions on 45 test items, for memory and for reasoning. These predictions were compared with the 16 key data points shown in Table 8, namely the average response rates on old, lure, new medium, and new small items, for memory and for reasoning, and for dogs only and for dogs + nondogs. The model was applied to 25 different random stimulus

configurations. The average estimated parameter values of A and B were 0.38 and 1.13. For the memory dogs only, memory dogs + nondogs, reasoning dogs only, and reasoning dogs + nondogs conditions, the respective *c* parameters were 4.08, 2.49, 3.38, and 1.93, and the respective β parameters were 0.45, 1.00, 0.92, and 1.26. The estimated A and B values indicated that medium dogs were closer in psychological space to the large dogs than were the small dogs. Notably, *c* was considerably lower for reasoning than for memory, reflecting broader generalization for reasoning and steeper generalization for memory. Likewise, as predicted, adding birds and fish during study led to decreases in *c* (i.e., broader generalization) for both reasoning and memory. We did not observe interpretable changes in β across conditions.

Overall, there was a good fit between model and data, with an average *RMSE* of .0347. Table 8 shows average predictions of the model. This table shows that the main trends in the data have been captured, including the weaker discrimination when nondogs were included in the test list. Likewise, the predicted *d′* measures are close to the original results. The simulation made 45 predictions for the memory and reasoning conditions, for 45 test items, so it was again possible to calculate the predicted correlation between memory and reasoning. In the simulations, the average predicted correlation between memory and reasoning was .97 in the dogs-only conditions and .94 in the dogs + nondogs conditions. Figure 7 shows a predicted trend line based on one run of the simulation.

Discussion

This study replicated the general pattern of similarities and differences between reasoning and memory found in previous studies, with induction associated with broader generalization of positive responses but a strong itemwise correlation between induction and recognition judgments about the common test stimuli. The addition of study items from other animal categories affected responses, decreasing sensitivity in the discrimination between old and new items. As predicted by GEN-EX, however, this effect was found under both induction and recognition conditions.

One objection to these conclusions might be that they rely on a comparison of conditions that differ not only in item content (i.e., dogs only vs. dogs + nondogs) but also study list length (15 study items vs. 45 study items). Indeed, in Experiment 2A as well, study list length varied between conditions (although not in Experiment 2B). In reasoning research, it is actually very common for problems to vary in length (Hayes et al., 2010; Heit, 2000; Rotello &
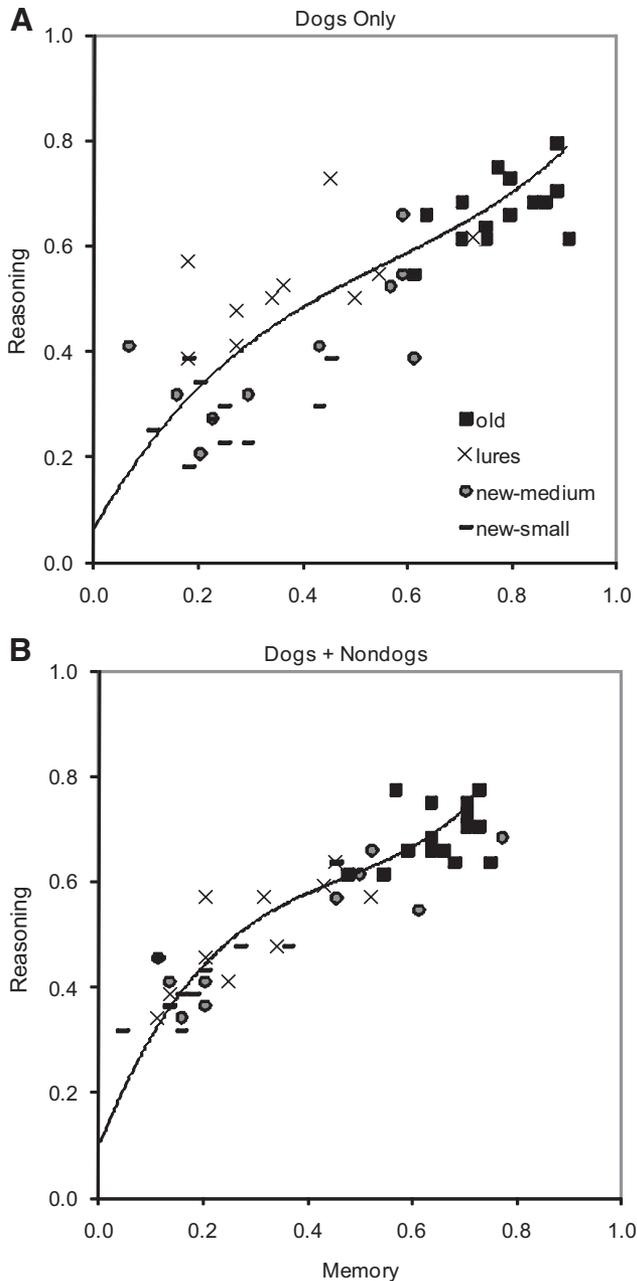
Figure 7. Scatter plots for Experiment 3, showing memory and reasoning responses across stimuli, for (A) dogs only and (B) dogs + nondogs, with trend lines of GEN-EX model predictions. GEN-EX = generalization from examples.

Heit, 2009), and often, there is little effort to keep length constant from one condition to another. In contrast, controlling for list length has been of greater concern in memory research. Some researchers have claimed that such manipulation of list-length affects recognition performance (Cary & Reder, 2003; Clark & Gronlund, 1996) with sensitivity in old–new discrimination lower following the study of long lists (but see Dennis & Chapman, 2009, for the opposite result). It is important to note, however, that

the validity of this claim has been questioned on both methodological and theoretical grounds (cf. Dennis & Humphreys, 2001; Dennis, Lee, & Kinnell, 2008; Murdock & Kahana, 1993; Underwood, 1978). Notably, Dennis and Humphreys (2001) have pointed out that the apparent effects of list length can be explained by differences in the retention intervals for items on shorter and longer lists, changes in attention to the later items in longer lists, and changes in the context in which items are retrieved. When these variables are controlled then no main effect of list length on recognition is found (Dennis & Humphreys, 2001; Dennis et al., 2008). Moreover, even those who argue for an independent effect of list length (e.g., Cary & Reder, 2003) suggest that at least a fourfold difference in length between short and long lists may be necessary to detect this effect. Because the discrepancy between the two study lists in this experiment was of a smaller magnitude and because there was no change in retrieval context between study and test phases in any condition, it seems unlikely the differences between the dogs and no-dogs condition were solely due to list length.

However, we would concede the point that study list length varied in Experiment 3, and to the extent that we found systematic differences between conditions (which the GEN-EX model was able to address), these differences may be partly attributable to study list length. Although it is not our primary goal to rule out study list length effects, but rather to examine how the exactly the same experimental manipulations affect memory and reasoning, in Experiment 4, this issue was addressed by holding study list length constant and introducing items from outside of the dog category in the test list.

Experiment 4

In Experiment 4, all participants studied the same list of 15 dogs. For half the participants, in the dogs + nondogs conditions, we introduced pictures of birds and fish in the test list, as another way to examine the effects of presenting extracategory items. Again, we predicted that presenting birds and fish at test would highlight the categorical nature of the stimuli, making discrimination among dogs somewhat poorer in both the memory and reasoning conditions, reflected in the c parameters of the GEN-EX model.

Method

Participants. One hundred and sixty-eight University of California, Merced, students participated. Equal numbers were randomly allocated to the four conditions generated by the factorial crossing of task (recognition vs. induction) and test list (dogs only vs. dogs + nondogs).

Procedure. The study phase procedure for the recognition and induction conditions was similar to Experiment 3, except that all participants saw 15 dogs during the study and no birds or fish. During the test phase, participants in the dogs-only condition were presented with the 15 study dogs, 10 new small dogs, 10 new medium dogs, and 10 large (lure) dogs in random order. Participants in the dogs + nondogs condition were presented with the same set of 45 test dogs together with 15 photographs of birds and 15 photographs of fish, with item order randomized.

Results

The probability of responding positively to dog test items under recognition and induction conditions is shown in Table 9. As in

Table 9
Experiment 4. Results (Proportion of "Yes" Responses and d′*) and Model Predictions*

Result and prediction	Old	New small	New medium	All new	Lure	d' (Old–new)	d' (Old–lure)
Result							
Dogs only memory	.77	.24	.34	.29	.37	1.30	1.08
Dogs only reasoning	.71	.32	.42	.37	.47	0.88	0.64
Dogs + nondogs memory	.71	.24	.35	.29	.40	1.10	0.80
Dogs + nondogs reasoning	.81	.55	.65	.60	.64	0.66	0.57
Predictions							
Dogs only Memory	.77	.24	.33	.29	.36	1.30	1.11
Dogs only reasoning	.71	.34	.42	.38	.45	0.86	0.69
Dogs + nondogs memory	.71	.26	.35	.31	.38	1.07	0.88
Dogs + nondogs reasoning	.81	.56	.63	.65	.64	0.64	0.50

previous studies, there were more positive responses to medium than small dogs, $F(1, 164) = 48.9$, $p < .001$, partial $\eta^2 = .23$, but this difference did not interact with either task or test list variables ($Fs < 0.5$), and responses to these stimuli were collapsed. Positive responding for old, new, and lure items was analyzed with factorial ANOVA with task (memory vs. induction) and test list (dogs only vs. dogs + nondogs) as between-subjects factors. The probability of responding "yes" to new items and lures was higher for induction than recognition, $F(1, 164) = 26.03$, $p < .001$, partial $\eta^2 = .14$; $F(1, 164) = 12.76$, $p < .001$, partial $\eta^2 = .07$, respectively. The probability of responding "yes" to new and lure items was higher when the test list included items from other categories than when only dogs were included, $F(1, 164) = 9.12$, $p < .01$, partial $\eta^2 = .05$; $F(1, 164) = 5.03$, $p < .05$, partial $\eta^2 = .03$, respectively. For positive responses to old items, there was also a significant interaction between task and test list, $F(1, 164) = 7.56$, $p < .01$, partial $\eta^2 = .04$. Tests of simple effects confirmed that rates of positive responding were higher for induction than recognition when dogs and nondogs were included at test, $t(82) = 2.02$, $p = .047$, but did not differ when only dogs were presented ($p = .22$). An interaction between task and test list was also found for positive responses to new items, $F(1, 164) = 6.93$, $p < .01$, partial $\eta^2 = .04$. Table 9 shows that the difference in positive responding between induction and recognition was larger in the dogs + nondogs than in the dogs-only condition. This was mainly due to a large increase in false alarms to new items when those doing reasoning were tested with the dogs + nondogs list.

This study also afforded an opportunity to examine inductive and recognition responses to test items in different animal categories. Although we did not have sharp predictions here, we observed that within the dogs + nondogs condition, the probability of responding "yes" to nondog items (birds or fish) was very low in the recognition condition ($M = 0.02$) and significantly higher in induction ($M = 0.24$), $F(1, 82) = 16.6$, $p < .01$, partial $\eta^2 = .17$. This result reflects participants' overall tendency, also observed for dog stimuli, to generalize to some stimuli in the induction task that they would not falsely recognize in the recognition task. More generally, the nontrivial level of positive responding to nondogs in the induction condition seems consistent with previous work showing that people often generalize biologically plausible properties (like "has beta cells") from typical animals (such as dogs) to other animals (Osherson et al., 1990).

As before, individual ds were calculated for discrimination between old and new items and between old and lure items. These

were analyzed with factorial analyses of variance with task and test list as between-subjects factors. Discrimination between old and new items was greater for recognition than induction, $F(1, 164) = 15.98$, $p < .001$, partial $\eta^2 = .09$, and greater for dogs only than dogs + nondogs, $F(1, 164) = 4.88$, $p < .05$, partial $\eta^2 = .03$. Old–lure discrimination sensitivity was also higher for recognition than induction, $F(1, 164) = 8.3$, $p < .01$, partial $\eta^2 = .05$, but there was no main effect of test list, $F(1, 164) = 2.75$, $p = .10$. Neither of these analyses found interactions between task and test list ($Fs < 0.5$).

Itemwise correlations between positive responding to dog items in induction and recognition were high and positive for both the dogs only condition, $r(43) = 0.92$, $p < .001$, and the dogs + nondogs condition, $r(43) = 0.87$, $p < .001$. (See Figure 8 for scatter plots.) In the latter condition, the itemwise correlation between positive responding to the 30 bird and fish test items in the induction and recognition conditions was also calculated. This correlation was low, $r(28) = 0.11$, and not significant ($p = .56$). Note though that this low correlation reflects a severely restricted range of positive responses (e.g., in recognition, each bird and fish test item received no more than 5% positive responses).

Modeling

As in Experiment 3, we again simulated the overall patterns of responses, based on randomized configurations of hypothetical stimuli. As in the previous study only the GEN-EX null model was applied, and we did not simulate the data for bird and fish items. (Doing so would be possible, but would likely add as many free parameters as it would distinguishable data points).

The model was used to make predictions on 45 test items, for memory and for reasoning. These predictions were compared with the 16 key data points shown in Table 9, namely the average response rates on old, lure, new medium, and new small items, for memory and reasoning, and for dogs only and dogs + nondogs. The c and β parameters were allowed to vary across the four conditions; hence, with the addition of the A and B parameters, there were 10 free parameters for 16 data points. The model was applied to 25 different random stimulus configurations. The average estimated parameter values of A and B were 0.41 and 0.64. For the memory, dogs only, memory, dogs + nondogs, reasoning, dogs only, and reasoning, dogs + nondogs, the respective c parameters were 4.87, 3.16, 3.99, and 2.29, and the respective β parameters were 0.38, 0.66, 0.57, and 0.49. The estimated A and B values reflect the fact
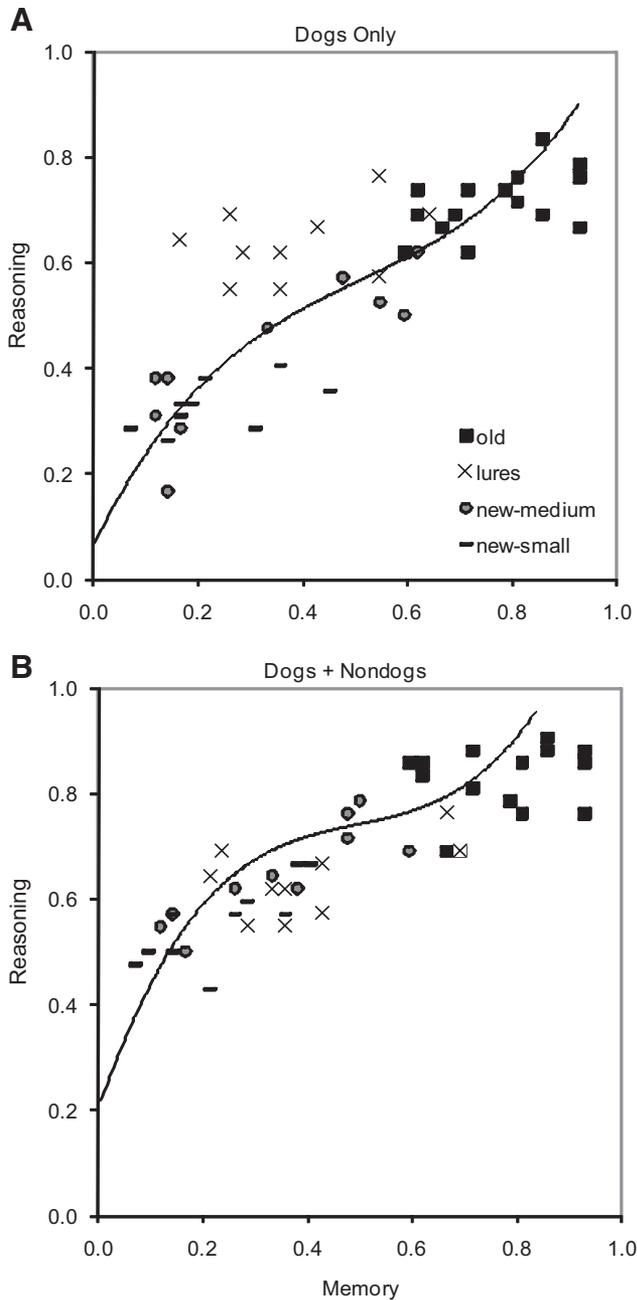
A

B

Figure 8. Scatter plots for Experiment 4, showing memory and reasoning responses across stimuli, for (A) dogs only and (B) dogs + nondogs, with trend lines of GEN-EX model predictions. GEN-EX = generalization from examples.

that the medium dogs were closer in psychological space to the large dogs than were the small dogs. Notably, c values were considerably lower for reasoning than for memory, reflecting broader generalization for reasoning and steeper generalization for memory. Likewise, adding birds and fish at test led to decreases in c for both reasoning and memory. We did not observe substantial or interpretable changes in β across conditions.

Overall, there was a good fit between model and data, with an average *RMSE* of .0306. Table 9 shows average predictions of the model. This table shows that the main trends in the data have been captured, including weaker discrimination when nondogs are included in the test list. Likewise the predicted d' measures are close to the original results. As in previous studies, it was possible to calculate the predicted correlation between memory and reasoning. The predicted correlations were .97 in the dogs-only conditions and .90 in the dogs + nondogs conditions (as compared with corresponding observed correlations of 0.92 and 0.87). Figure 8 shows a predicted trend line based on one run of the simulation; note again that the model predictions in Table 9 were based on averages from 25 random configurations.

Discussion

Adding test items from other basic-level categories did affect patterns of positive responding in recognition and induction, decreasing sensitivity in discrimination between old and new items. We also replicated the general pattern of differences between test phase responding for recognition and induction found in previous experiments. Induction was again associated with a broader pattern of generalization to novel test items than recognition. Notably, these effects of task and manipulation of the test list on test phase sensitivity were independent. Although adding bird and fish items increased positive responding in old and new items in induction relative to recognition, this did not translate into significant task by list interactions on the sensitivity measures. Moreover, the positive empirical relation between induction and recognition for dog items remained strong when items from a variety of basic animal categories were added to the test list.

Experiment 4 was complementary to Experiment 3, in that we held study list length constant and varied test list length. We think it is unlikely that the results of Experiment 4 are solely due to test list length and not the presence of additional categories, but we acknowledge that some role of test list length, in itself, cannot be ruled out. What is important for our purposes is that the manipulation had the same effect on both memory and reasoning and could be accommodated by the GEN-EX model mainly by varying the sensitivity parameters.

In short, adding items from other basic level categories affected the sensitivity of test judgments in induction and recognition in similar ways. Consistent with the predictions of GEN-EX, the empirical relation between positive responding to dog items in induction and recognition remained strong, even when judgments also had to be made about the members of other categories.

General Discussion

Our main aim was to examine the relation between recognition memory and inductive reasoning judgments about a common set of category exemplars. The main empirical finding was that in every experiment, when procedural differences between these tasks were kept to a minimum, there was a close correspondence between the two kinds of judgments. Judgments about whether a novel property would generalize from a set of familiar instances to a given test item were predictable from the probability of responding "old" to that item in recognition, with an average correlation across experiments of .87 and similar patterns of responses on old, lure,

medium, small, and out-of-category items. The strength of this relation supports the view that recognition and induction share underlying component processes.

There were also some interesting differences between memory and reasoning. In every study (except Experiment 1D), we found that people doing induction were more likely to make positive responses to novel test stimuli (and hence showed lower sensitivity) than those doing recognition. In other words, induction was generally associated with a broader generalization of positive responding to novel items than recognition. This had been predicted, due to the different natures of the two tasks, with recognition emphasizing rejection of novel stimuli and induction emphasizing generalization to novel stimuli.

There was also suggestive evidence of different roles for deterministic responding in memory versus reasoning conditions. In particular, in a number of studies (Experiments 1A–1D, 2A, 2B) we found that adding a parameter reflecting deterministic responding improved the fit of the GEN-EX model, especially for recognition. In contrast, adding a deterministic parameter did not generally improve the fit of GEN-EX to induction data.

Another important finding was that a range of changes to the conditions under which people encoded study items or made decisions about test instances had parallel effects on reasoning and memory. Repeating study instances led to better discrimination between old and lure items in recognition and had a very similar effect on induction. Adding test items from outside of the study category broadened generalization of positive test responses for both induction and recognition. Our interpretation of these findings is that these study and test variables principally affected the way that similarity between old and test items was computed. Because this is a common core process in induction and recognition, these manipulations produced similar effects in each task.

This account is supported by the modeling results in each study. With the possible exception of Experiment 1D, reasoning and memory judgments in each study could be accounted for by a single model in which positive test responses were determined by the total similarity between a test item and previously studied items. The most important differences between recognition and induction in test phase "yes" responding were captured by changes in a sensitivity parameter that increased or decreased the overall psychological distance between study and test stimuli. Similarly, the effects of manipulations like study phase frequency and test item length were reflected in changes in the GEN-EX similarity and generalization parameters. Comparative modeling suggested some differences between conditions for subcategory-based and deterministic responding, but in general, the similarities between the effects of the manipulations on induction and recognition far outweighed the differences.

The most straightforward implication of these findings is that inductive reasoning and memory are not as different as has often been assumed. The current data suggest that exemplar similarity plays a major role in both object recognition and in the generalization of novel properties between category exemplars. The observed differences in patterns of recognition and induction were explained in a relatively straightforward way by altering similarity parameters in the respective decision rules. Notably, the data from these two tasks could be explained within a single processing framework. There was no need to require the postulation of mul-

tiple systems of representation or processing, or even different sources of information affecting the two tasks.

Although similarity has been long been acknowledged as a key component in models of induction (e.g., Osherson et al., 1990) and recognition (e.g., Hintzman, 1988), ours is the first attempt to explain recognition and induction judgments for a common stimulus set with a single theoretical model. Moreover, previous models of induction (e.g., Osherson et al., 1990) have usually defined similarity in terms of the relations between different categories (e.g., lions, horses, mammals). The current work makes a novel contribution by highlighting the importance of similarity between specific exemplars in property induction.

Comparison Between GEN-EX and Other Induction Models

The key findings in the current experiments were that (a) induction usually leads to a broader generalization of positive responses to novel test items than recognition, (b) there was a strong positive correlation between the probability of making a positive response to test items under induction and recognition instructions, and (c) a range of variables that affect the similarity between test items and old items retrieved from memory had parallel effects on induction and recognition. These data were well explained by the GEN-EX model, which assumes that both induction and recognition responses are influenced by the total similarity of a novel item to previously experienced old items but that induction is associated with a broader gradient of stimulus generalization.

Although previous models of induction have not addressed recognition, it is reasonable to ask how such models might approach the induction–recognition relation (see Heit & Hayes, 2005, for a related discussion). Perhaps the approach that most closely resembles our own is the SINC (similarity, induction, and categorization) model developed by Sloutsky and Fisher (2004). This model was developed to explain how item similarity determines children's categorization and induction judgments. Sloutsky and Fisher (2004) examined relations between induction and recognition by presenting adults and children with animal stimuli (cats, birds, fish), some of whom (the cats) had a novel anatomical property, and then administering a surprise recognition test in which old and new cats were shown together with new animal foils. Adults often mistakenly responded positively to novel cats during recognition. It was claimed that children were less likely to false alarm to these stimuli, but this claim has since been shown to be a consequence of the fact that children tended to look longer at study pictures than adults (Hayes, McKinnon, & Sweller, 2008). When encoding times were kept relatively short and were equalized across age groups, similar false alarm rates for novel category members were found in adults and children (Hayes et al., 2008, also see Wilburn & Feeney, 2008).

Although SINC was never applied to the Sloutsky and Fisher (2004) recognition data we have shown elsewhere (Heit & Hayes, 2005), the model can be recast to generate recognition predictions by adding a component that assesses the total similarity of novel items to familiar items. The GEN-EX model, however, extends this idea in at least two important ways. First, the computation of similarity has been refined to make it more consistent with well-supported models of categorization such as GCM. Second, GEN-EX explicitly assumes that recognition and induction judg-

ments differ in the breadth of generalization around study items. The latter assumption predicts Sloutsky and Fisher's finding that adults are more likely to false alarm in a recognition test that follows an induction study task, as opposed to a recognition study task. Another important difference between SINC and GEN-EX is that the former model was proposed only to explain induction in young children, whereas GEN-EX is proposed as a general account of induction and recognition. We have shown that GEN-EX gives a good account of adult responses on both tasks. Although it has yet to be applied to children's judgments, there are good reasons to believe that the processes underlying induction and recognition have considerable developmental continuity (Hayes et al., 2008) and that therefore GEN-EX may also give a good account of the induction–recognition relation in children.

As the name implies, Osherson et al.'s (1990) similarity-coverage model accords a central role to the similarity between inductive base and target items in judgments about property generalization. Although this model was not intended as an account of recognition, it is interesting to consider this possible extension. One problem with extrapolating this model is that it conceives of similarity largely in terms of similarity between basic or superordinate categories rather than similarity between specific exemplars. This leads to improbable predictions about false alarms to specific instances. According to the coverage component of the model, if items from a variety of categories are presented at study, as in Sloutsky and Fisher (2004), then a superordinate that encompasses these items will be generated, and similarity to items within the superordinate will drive inferences about test items. Hence, adults who saw cats, birds, and bears at study would be predicted to strongly false alarm in a recognition test to animal lures such as squirrels or fish, which also belong to the animal superordinate.

Sloman (1993) proposed an alternative, connectionist model of induction, relying on feature overlap between premise and conclusion categories to make predictions about argument strength. It would be natural to extend this approach to making recognition judgments in terms of feature overlap. Most likely, for the model to be used as a viable account of list recognition memory, hidden units would need to be introduced to the model. There would be parallels to our own, similarity-based approach, with a key difference being that GEN-EX uses an exemplar-based representation, whereas Sloman's model has a distributed, featural representation.

More recently, Bayesian accounts of inductive reasoning have been proposed (Heit, 1998a; Kemp & Tenenbaum, 2009; Tenenbaum & Griffiths, 2001). These models use probability rather than similarity as a basis for making predictions but still can account for similarity effects in reasoning by assuming that hypothesis spaces are correlated with similarity structure. For example, observing a set of large dogs with beta cells would support the hypothesis that only large dogs have beta cells and promote generalization to other large dogs more so than generalization to medium or small dogs. It remains to be seen whether these models of induction could be extended to recognition tasks. We note, however, that there are already Bayesian models of recognition memory in existence (e.g., Criss & McClelland, 2006; Shiffrin & Steyvers, 1997).

Dual-Process Accounts of Memory and Reasoning

Our research and the GEN-EX model itself highlight a commonality between memory and inductive reasoning, namely that they are both heavily influenced by comparisons between the features of familiar and novel exemplars. However, we do not dispute the notion that other processes contribute to memory and reasoning. In memory research, there is evidence (e.g., Yonelinas, 2002) for a second, slower, more controlled, recollective process that also contributes to recognition, and dual-process models of recognition have been implemented successfully and applied to data (e.g., Rotello, Macmillan, & Reeder, 2004). Likewise, in reasoning research, there is evidence that causal reasoning (e.g., Rehder & Burnett, 2005) goes beyond the effects of similarity on inductive reasoning and that more generally, analytic processing supplements heuristic processing (e.g., Evans, 2008). Dual-process models of reasoning that suggest different processing principles for induction and for logical deduction have also been implemented and had some success in explaining empirical data (e.g., Heit & Rotello, 2010; Rotello & Heit, 2009).

Just as we have found links between similarity-based processing in memory and reasoning, an intriguing question for future research is whether there are also links between controlled processes in memory and reasoning. Indeed, the subcategory-based and deterministic responding processes could be considered as kinds of controlled processing. It is interesting that recollective judgments have sometimes been described as a kind of logical reasoning. For example, the recall-to-reject process (Rotello & Heit, 1999) could be used to determine whether a tested item had really been studied or whether a similar item had been studied instead. Suppose that a participant studies the word *chair* and is tested on the word *chairs*. The participant knows that at most one of these words was studied. Although *chairs* will seem familiar, if the participant can recollect studying the word *chair* then it will be possible to reject the word *chairs*, using the recall-to-reject process, which here corresponds to the *modus ponendo tollens* inference in classical logic, which has the form *not (A and B); A; therefore not B*.

Relation to Other Work on Reasoning and Memory

To the best of our knowledge, the current studies represent the first attempt to examine the relation between recognition and property induction. However, there has been considerable investigation of the relations between memory and other forms of reasoning. One line of work that has some overlap with the current studies involves the examination of links between memory and probabilistic judgments. One of the best known phenomena suggesting such a link is the availability bias, whereby judgments of the relative frequency or probability of events are influenced by the ease with which instances of those events can be recalled (e.g., Lichtenstein, Slovic, Fischoff, Layman, & Combs, 1978; Tversky & Kahneman, 1973). To explain such phenomena, Hastie and Park (1986) argued that when probability judgments are based on event attributes that have to be retrieved from memory (as opposed to examined on-line), there will be a strong positive relation between memory performance and judgment outcome. Hence, when people were asked to listen to a recorded conversation and subsequently judge the suitability of one of the speakers for a job as a computer programmer, there was a strong positive correlation between judgments and recall of conversation statements favoring or opposing the appointment. This correlation disappeared, however, when the task was run "online" with participants told about the required judgment before listening to the conversation.

Many aspects of the current studies are consistent with Hastie and Park's (1986) account. In most of our studies, inductive and recognition judgments were made about previously encoded stimuli. Under these retrospective decision conditions, we found the strong positive relation between performance on both tasks predicted by Hastie and Park. The results of Experiment 1D, however, in which the memory component was removed from induction, only partly support Hastie and Park (1986). In Experiment 1D, inductive inferences were effectively "on-line" during test. This did alter the relative patterns of performance in induction and recognition (e.g., by increasing the hit rate for induction). Critically, though, the relation between positive responses at test in induction and recognition remained strong. This suggests that Hastie and Park (1986) underestimated the degree of overlap between the processes underlying probabilistic reasoning and memory. Although the presence of a common memory component may affect the relation between these tasks, when this component is removed from induction, the common process of computing exemplar similarity remains, giving rise to strong correlations between recognition and induction performance.

Another important body of evidence examining links between memory and reasoning comes from studies of fuzzy trace theory (FTT; Brainerd & Reyna, 1993, 2004). One of the key motivations for FTT is to provide an account of memory-reasoning relations and how they develop. According to FTT, novel information is encoded in two parallel formats. Verbatim representations are precise and complete records of studied stimuli that include their surface perceptual details. Gist representations involve more abstract summaries of the semantic or relational content of presented information. An important assumption of FTT is that verbatim traces are usually accessed in memory tasks, whereas gist representations are usually used for reasoning. This leads to the prediction that memory for the literal details of inputs into reasoning (e.g., the initial premises in a reasoning task) can be statistically independent of reasoning performance (see Brainerd & Reyna, 1993, 2004, for relevant reviews).

Clearly, our results seem inconsistent with this prediction. Experiment 1C showed a high correlation between recognition and inductive judgments for the same items and showed that the probability of making a positive induction response increased when that item was recognized. More generally, we have shown that when general task factors are matched, a single process of exemplar similarity, albeit with flexible levels of generalization around exemplars, can explain both memory and inductive reasoning. In other words, we find that a single kind of representation goes a considerable way toward explaining performance on both memory and inductive reasoning tasks. However, we acknowledge that we have not applied FTT to our results.

It is also important to note that much of the previous evidence for stochastic independence between memory and reasoning in FTT (e.g., Brainerd & Kingma, 1985) has involved tasks that focus on deductive reasoning (e.g., transitive inference, class inclusion, conservation). As noted earlier, there is now good evidence that different processing principles are used in inductive and deductive reasoning (e.g., Heit & Rotello, 2010; Rotello & Heit, 2009). Hence, much of this previous work may not be generalizable to the current studies. FTT may indeed be correct in suggesting that the representations of logical relations in deductive arguments are functionally separate from verbatim memory for the argument

premises. In contrast, a large body of evidence (e.g., Goldstone, 1994; Sloutsky & Fisher, 2004) suggests that similarity in surface appearance is an important cue for category membership. Hence, in category-based induction, it seems reasonable to expect a close relation between the ability to detect and remember the similarities between category exemplars (e.g., the study items in the current experiments) and the way that exemplar properties are generalized to novel instances. Similar arguments may be used to explain the apparent discrepancy between our finding of a positive relation between induction and recognition and previous work showing that memory for examples can be dissociated from performance in analogical reasoning tasks (e.g., Gentner, Rattermann, & Forbus, 1993; Holyoak & Koh, 1987; Ross, 1987).

Induction and More Complex Forms of Similarity

Central to GEN-EX is the idea that similarity between known and novel exemplars affects the generalization of identity (in recognition) and shared properties (in induction). However, it has been pointed out that the psychological construct of similarity is unconstrained such that judgments about the similarity between two or more objects can vary substantially across different judgment contexts (e.g., Murphy & Medin, 1985). This is well illustrated in inductive reasoning studies in which varying the nature of the property can strengthen or weaken property generalization between the same target and test items (e.g., Heit & Rubinstein, 1994; Ross & Murphy, 1999; Shafto & Coley, 2003). Heit and Rubinstein (1994), for example, found that anatomical properties were more likely to be generalized from *sparrows* to *hawks* than from *tigers* to *hawks* but that this pattern reversed when the property was predatory behavior. The different properties appeared to cause people to compute similarity between target and test instances in different ways (e.g., taxonomic similarity vs. similarity between ecological roles).

One implication is that if more meaningful or familiar properties were used during the study phase of our induction task, this could change the way people make inductive judgments about new test dogs (using a stronger manipulation of property type than we attempted for Experiments 1A and 1B). For example, if during the induction study phase people were asked to learn about which dogs are good pets, it seems likely that people would make inductive judgments at test based on dimensions other than perceptual and taxonomic similarity (e.g., perceived temperament, attractiveness). This means that patterns of induction judgments should diverge from those made under recognition conditions.

This possible selective effect of property knowledge is an interesting target for further research. However, it is important to remember that the goal of the current studies was to examine the relations between induction and recognition when those doing induction had minimal knowledge about the property being generalized; that is, when they could only rely on some default notion of similarity between study and test exemplars. The key finding was that the same default form of exemplar similarity operated for induction and recognition. In this respect, our general strategy was similar to previous work on category-based induction (e.g., Osherson et al., 1990) in which unfamiliar properties were used so that the effects of taxonomic similarity on inductive judgments could be isolated and studied in detail.

Another way that our work goes beyond simple similarity-based responding is the incorporation of two kinds of deterministic or rule-based processes, in terms of subtypes as well as veridical recollection of items. In general, our modeling found evidence for the processes in addition to the influence of overall similarity.

Conclusion

The current work makes two novel contributions to the study of inductive reasoning. First, it formalizes the process whereby properties are generalized between exemplars of the same category. Second and more significantly, it shows that a similar generalization process operates when people make inductive and recognition judgments. This link between inductive reasoning and memory offers a number of interesting prospects for further research. There are a range of empirical factors known to affect recognition performance, which may also have an impact on induction (see Yonelinas, 2002, for a review). Moreover, GEN-EX's success in explaining both recognition and induction data represents a first step toward the development of a general theory of how people generalize knowledge between familiar and unfamiliar instances.

References

American Kennel Club. (2006). AKC recognized breeds. Retrieved from http://www.akc.org/breeds/breeds_a.cfm

Borowiak, D. S. (1989). *Model discrimination for nonlinear regression models.* New York, NY: Marcel Dekker.

Bott, L., & Murphy, G. L. (2007). Subtyping as a knowledge preservation strategy in category learning. *Memory & Cognition, 35,* 432–443.

Brainerd, C. J., & Kingma, J. (1985). On the independence of short-term memory and working memory in cognitive development. *Cognitive Psychology, 17,* 210–247. doi:10.1016/0010-0285(85)90008-8

Brainerd, C. J., & Reyna, V. F. (1993). Memory independence and memory interference in cognitive development. *Psychological Review, 100,* 42–67. doi:10.1037/0033-295X.100.1.42

Brainerd, C. J., & Reyna, V. F. (2004). Fuzzy-trace theory and memory development. *Developmental Review, 24,* 396–439. doi:10.1016/j.dr.2004.08.005

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language, 49,* 231–248. doi:10.1016/S0749-596X(03)00061-5

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review, 3,* 37–60.

Criss, A. H. (2009). The distribution of subjective memory strength: List strength and response bias. *Cognitive Psychology, 59,* 297–319. doi:10.1016/j.cogpsych.2009.07.003

Criss, A. H., & McClelland, J. L. (2006). Differentiating the differentiation models: A comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLiM). *Journal of Memory and Language, 55,* 447–460. doi:10.1016/j.jml.2006.06.003

Dennis, S., & Chapman, A. (2009). The inverse list length effect: Implications for separate storage models of recognition memory. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the Thirty-First Conference of the Cognitive Science Society,* 2385–2390.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108,* 452–478. doi:10.1037/0033-295X.108.2.452

Dennis, S., Lee, M. D., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list length effect. *Journal of Memory and Language, 59,* 361–376. doi:10.1016/j.jml.2008.06.007

Diana, R. A., Reder, L. M., Arndt, J., & Park, H. (2006). Models of

recognition: A review of arguments in favor of a dual-process account. *Psychonomic Bulletin & Review, 13,* 1–21.

Estes, W. K. (1994). *Classification and cognition.* Oxford, England: Oxford University Press. doi:10.1093/acprof:oso/9780195073355.001.0001

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology, 59,* 255–278. doi:10.1146/annurev.psych.59.103006.093629

Feeney, A. (2007). How many processes underlie category-based induction? Effects of conclusion specificity and cognitive ability. *Psychonomic Bulletin & Review, 14,* 884–889.

Fisher, R. (1951). *The design of experiments.* New York, NY: Oxford University Press.

Flexser, A. J., & Bower, G. H. (1974). How frequency affects recency judgments: A model for recency discrimination. *Journal of Experimental Psychology, 103,* 706–716. doi:10.1037/h0037194

Gentner, D., Rattermann, M., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology, 25,* 524–575. doi:10.1006/cogp.1993.1013

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 500–513. doi:10.1037/0278-7393.25.2.500

Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition, 52,* 125–157. doi:10.1016/0010-0277(94)90065-5

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological Review, 93,* 258–268. doi:10.1037/0033-295X.93.3.258

Hayes, B. K., Foster, K., & Gadd, N. (2003). Prior knowledge and subtyping effects in children's category learning. *Cognition, 88,* 171–199. doi:10.1016/S0010-0277(03)00021-0

Hayes, B. K., & Heit, E. (2004). Why learning and development can lead to poorer recognition memory. *Trends in Cognitive Sciences, 8,* 337–339. doi:10.1016/j.tics.2004.05.001

Hayes, B. K., Heit, E., & Swendsen, H. (2010). Inductive reasoning. *Wiley Interdisciplinary Reviews: Cognitive Science, 1,* 278–292.

Hayes, B. K., McKinnon, R., & Sweller, N. (2008). The development of category-based induction: Reexamining conclusions from the induction then recognition paradigm. *Developmental Psychology, 44,* 1430–1441. doi:10.1037/0012-1649.44.5.1430

Heit, E. (1992). Categorization using chains of examples. *Cognitive Psychology, 24,* 341–380. doi:10.1016/0010-0285(92)90011-P

Heit, E. (1994). Models of the effects of prior knowledge on category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1264–1282. doi:10.1037/0278-7393.20.6.1264

Heit, E. (1998a). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational Models of Cognition* (pp. 248–274). Oxford, England: Oxford University Press.

Heit, E. (1998b). Influences of prior knowledge on selective weighting of category members. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 712–731. doi:10.1037/0278-7393.24.3.712

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review, 7,* 569–592.

Heit, E. (2001). Putting together prior knowledge, verbal arguments, and observations in category learning. *Memory & Cognition, 29,* 828–837.

Heit, E. (2007). What is induction and why study it? In A. Feeney & E. Heit (Eds.), *Inductive reasoning* (pp. 1–24). Cambridge, England: Cambridge University Press.

Heit, E., Brockdorff, N., & Lamberts, K. (2003). Adaptive changes of response criterion in recognition memory. *Psychonomic Bulletin & Review, 10,* 718–723.

Heit, E., & Hayes, B. K. (2005). Relations between categorization, induction, recognition and similarity. *Journal of Experimental Psychology: General, 134,* 596–605. doi:10.1037/0096-3445.134.4.596

Heit, E., & Hayes, B. K. (2008). Predicting reasoning from visual memory. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1831–1836). Austin, TX: Cognitive Science Society.

Heit, E., & Rotello, C. M. (2010). Relations between inductive reasoning and deductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 805–812. doi:10.1037/a0018784

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 411–422. doi:10.1037/0278-7393.20.2.411

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95,* 528–551. doi:10.1037/0033-295X.95.4.528

Hintzman, D. L. (2001). Judgments of frequency and recency: How they relate to reports of subjective awareness. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1347–1358. doi:10.1037/0278-7393.27.6.1347

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition, 15,* 332–340.

Jones, C. M., & Heit, E. (1993). An evaluation of the total similarity principle: Effects of similarity on frequency judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 799–812. doi:10.1037/0278-7393.19.4.799

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review, 116,* 20–58. doi:10.1037/a0014282

Lamberts, K. (1994). Flexible tuning of similarity in exemplar-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1003–1021. doi:10.1037/0278-7393.20.5.1003

Lamberts, K. (2002). Feature sampling in categorization and recognition of objects. *Quarterly Journal of Experimental Psychology: Section A. Human Experimental Psychology, 55,* 141–154. doi:10.1080/02724980143000208

Lee, H. S., & Holyoak, K. J. (2008). Causal models in analogical inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1111–1122. doi:10.1037/a0012581

Lichtenstein, S., Slovic, P., Fischoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 551–578. doi:10.1037/0278-7393.4.6.551

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false alarm rates. *Psychological Bulletin, 98,* 185–199. doi:10.1037/0033-2909.98.1.185

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review, 10,* 517–532.

Metcalfe, J., & Fisher, R. M. (1986). The relation between recognition memory and classification learning. *Memory & Cognition, 14,* 164–173.

Murdock, B. B., & Kahana, M. J. (1993). Analysis of the list-strength effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 689–697. doi:10.1037/0278-7393.19.3.689

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316. doi:10.1037/0033-295X.92.3.289

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57. doi:10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1988a). Exemplar-based accounts of relations between classification recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 700–708. doi:10.1037/0278-7393.14.4.700

Nosofsky, R. M. (1988b). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 54–65. doi:10.1037/0278-7393.14.1.54

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 3–27. doi:10.1037/0096-1523.17.1.3

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. doi:10.1111/1467-9280.00051

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97,* 185–200. doi:10.1037/0033-295X.97.2.185

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review, 97,* 285–308. doi:10.1037/0033-295X.97.2.285

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). The list-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 163–178. doi:10.1037/0278-7393.16.2.163

Rehder, B., & Burnett, R. (2005). Feature inference and the causal structure of categories, *Cognitive Psychology, 50,* 264–314. doi:10.1016/j.cogpsych.2004.09.002

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 629–639. doi:10.1037/0278-7393.13.4.629

Ross, B. H., & Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology, 38,* 495–553. doi:10.1006/cogp.1998.0712

Rotello, C. M., & Heit, E. (1999). Two-process models of recognition memory: Evidence for Recall-to-reject? *Journal of Memory and Language, 40,* 432–453. doi:10.1006/jmla.1998.2623

Rotello, C. M., & Heit, E. (2000). Associative recognition: A case of recall-to-reject processing. *Memory & Cognition, 28,* 907–922.

Rotello, C. M., & Heit, E. (2009). Modeling the effects of argument length and validity on inductive and deductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1317–1330. doi:10.1037/a0016648

Rotello, C. M., Macmillan, N. A., & Reeder, J. A. (2004). Sum-difference theory of remembering and knowing: A two-dimensional signal detection model. *Psychological Review, 111,* 588–616. doi:10.1037/0033-295X.111.3.588

Shafto, P., & Coley, J. D. (2003). Development of categorization and reasoning in the natural world: Novices to experts, naive similarity to ecological knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 641–649. doi:10.1037/0278-7393.29.4.641

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237,* 1317–1323. doi:10.1126/science.3629243

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

Shin, H. J., & Nosofsky, R. M. (1992). Similarity-scaling studies of dot-pattern classification and recognition. *Journal of Experimental Psychology: General, 121,* 278–304. doi:10.1037/0096-3445.121.3.278

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology, 25,* 231–280. doi:10.1006/cogp.1993.1006

Sloman, S. A., & Lagnado, D. A. (2005). The problem of induction. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 95–116). New York, NY: Cambridge University Press.

Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General, 133,* 166–188. doi:10.1037/0096-3445.133.2.166

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24,* 629–640. doi:10.1017/S0140525X01000061

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327–352. doi:10.1037/0033-295X.84.4.327

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5,* 207–232. doi:10.1016/0010-0285(73)90033-9

Underwood, B. J. (1978). Recognition memory as a function of the length of study list. *Bulletin of the Psychonomic Society, 12,* 89–91.

Wilburn, C., & Feeney, A. (2008). Do development and learning really decrease memory? On similarity and category-based induction in adults and children. *Cognition, 106,* 1451–1464. doi:10.1016/j.cognition.2007.04.018

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114,* 152–176. doi:10.1037/0033-295X.114.1.152

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114,* 245–272. doi:10.1037/0033-295X.114.2.245

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language, 39,* 124–148. doi:10.1006/jmla.1998.2566

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46,* 441–517. doi:10.1006/jmla.2002.2864

ORDER FORM

Start my 2011 subscription to the *Journal of Experimental Psychology: General* ISSN: 0096-3445

___	$55.00	**APA MEMBER/AFFILIATE**
___	$107.00	**INDIVIDUAL NONMEMBER**
___	$388.00	**INSTITUTION**
		In DC and MD add 6% sales tax
		TOTAL AMOUNT DUE $ ___

Subscription orders must be prepaid. Subscriptions are on a calendar year basis only. Allow 4-6 weeks for delivery of the first issue. Call for international subscription rates.

SEND THIS ORDER FORM TO
American Psychological Association
Subscriptions
750 First Street, NE
Washington, DC 20002-4242

Call **800-374-2721** or 202-336-5600
Fax **202-336-5568** : TDD/TTY **202-336-6123**
For subscription information,
e-mail: **subscriptions@apa.org**

AMERICAN
PSYCHOLOGICAL
ASSOCIATION

❏ **Check enclosed** (make payable to APA)

Charge my: ❏ Visa ❏ MasterCard ❏ American Express

Cardholder Name _____

Card No. _____ Exp. Date _____

_____
Signature *(Required for Charge)*

Billing Address
Street _____
City _____ State _____ Zip _____
Daytime Phone _____
E-mail _____

Mail To
Name _____
Address _____

_____
City _____ State _____ Zip _____
APA Member # _____

XGEA11