KNOWLEDGE SELECTION IN CATEGORY LEARNING

Evan Heit
Lewis Bott

I. Introduction

In our ordinary experience, we make countless observations every hour, with no observation perfectly resembling a previous case. We face a daily parade of unique events. Every time we walk into a building, for example, the building is unlike any other building in many ways. Even one particular building itself would be constantly undergoing various small changes. It has been suggested that to make better use of past experiences and simplify the processing that would be required for so many unique events, we learn about equivalence classes or categories of observations (e.g., Markman, 1989). For example, rather than treating every built architectural structure as being a unique construction, we form equivalence classes such as houses, office buildings, libraries, theaters, and pubs. These classes would facilitate many activities such as reasoning and communication. For example, just knowing that some building is a house would help to make predictions about its organization and layout, as well as help describe it to someone else. Categories allow us to greatly reduce the number of separate items we need to consider.

Although at first glance, categorization would seem to simplify our lives, it has been pointed out that category formation itself entails further complexities. Medin and Ross (1997) noted that just 10 objects can be parti-

THE PSYCHOLOGY OF LEARNING
AND MOTIVATION, VOL. 39

0079-7421/00 $30.00

tioned into categories more than 100,000 different ways. The implication is that whatever the benefits of forming categories, category learning itself is a difficult task that has costs in terms of processing and possibly in terms of getting the categorization wrong. So, in addressing one computational problem, the high number of unique events, we are led to another computational problem, the high number of possible partitions of events. As a solution to *this* problem, it has been proposed that, by necessity, category learning is not entirely data driven (e.g., Keil, 1989; Murphy & Medin, 1985; Peirce, 1931–1935). That is, people do not consider, and cannot consider, all observations and all possible partitions of observations when forming a category representation. Instead, category learning is constrained by inductive biases such as background knowledge. By use of theoretical knowledge about the world and knowledge of past categories, we are guided and directed when learning about new categories so that we do not have to consider all possible organizations of the observations we make.

In addition to this theoretical argument for the necessity of the use of background knowledge, it has by now been well established empirically that background knowledge has robust effects on category learning (see Heit, 1997, and Murphy, 1993, for wider reviews). There are a number of ways that background knowledge and observations are put together to learn new concepts, and a number of ways that background knowledge, observations, and concepts all interact during the course of category learning. Some of these ways are illustrated in Fig. 1. In this (extremely idealized) illustration, observations of some event or object are agglomerated in some fashion to learn a concept or representation of a category. However, there is much more going on, in that background knowledge also has several roles. These various points of contact between knowledge, observations, and concepts depicted in Fig. 1 are listed in Table I. This list does not represent a chronological order, but to some extent could be treated as being in order of difficulty or complexity. That is, any current account of category learning would address point A, some models might address points B or C to some extent, but few if any accounts would address points D and E. Point A simply refers to data-driven category learning, whereas points B and C refer to different influences of knowledge on the concept to be learned and on the use of observations. Points D and E refer to a contraflow of information in which the learned concept and observations are used to update or select from background knowledge.

A. Observations Used as Input to a New Concept

All accounts of category learning must have some way for observations to be accumulated in the form of a conceptual representation, correspond-

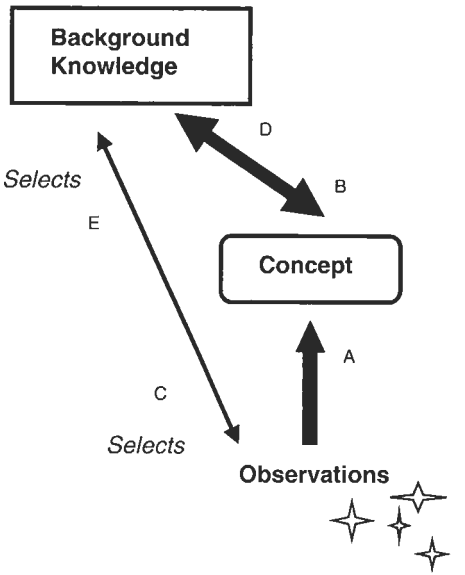

Fig. 1. Schematic illustration of several interactions among knowledge, concepts, and observations.

ing to point A. There are various ways for observations to be pooled together, such as being stored in the form of rules (e.g., Nosofsky, Palmeri, & McKinley, 1994), connection strengths (e.g., Gluck & Bower, 1988), other abstract representations (e.g., Ashby & Gott, 1988), or actually preserving memory traces for individual exemplars (e.g., Medin & Schaffer, 1978). Together, various models such as these have been applied successfully to

TABLE I

SEVERAL INTERACTIONS AMONG KNOWLEDGE, CONCEPTS, AND OBSERVATIONS

A.	Observations are used as a source of input to the new concept. That is, the content of the observations is reflected in the representation of the concept.
B.	Background knowledge is used as a source of input to the new concept. This is, the content of background knowledge is reflected in the representation of the concept.
C.	Background knowledge is used to select observations or particular features of dimensions.
D.	Background knowledge is updated, reflecting what is learned about the new concept.
E.	Observations are used to select background knowledge that will be helpful in learning about a new concept.

many hundreds of sets of experimental data on category learning. At this point, we will not try to distinguish among these various approaches to modeling and representing concepts. Instead, we simply note that point A is widely accepted and there is a great deal of relevant evidence.

B. Background Knowledge Used as Input to a New Concept

The next point covered is that background knowledge also provides a source of information to a newly learned concept. Although there could be a number of ways in which prior knowledge serves as an input to a new concept, perhaps the most straightforward is that information from background knowledge would just be copied into the new concept. For example, if you were to observe a novel species of bird, you would make some assumptions about ways that it is like other birds. You would initially assume that it was hatched from an egg and that it eats with its beak, even if you had not made these observations directly. Information derived from background knowledge about birds would be transferred into the new conceptual representation and put together with whatever is observed directly. In a program of research, Heit (1994, 1995, 1998a) referred to this putting together of background knowledge and observations as an *integration process.*

To be concrete, consider the prior knowledge effects reported by Heit (1994). These experiments simulated—in schematic form—the experience of visiting a new city and observing people there. This situation is a quintessential example of prior knowledge influencing category learning and categorization because there are completely novel categories of people to be learned but there is also extensive prior knowledge that is relevant, such as stereotypes of people in other places. First, subjects saw training examples consisting of featural descriptions of people in a novel city in an observational learning procedure. For example, one person might be described as shy and attending parties often. In effect, subjects were learning about contextualized categories, such as shy people in the new city and happy people in the new city. Then subjects were asked to make transfer judgments about additional people from the new city. For example, subjects were asked to judge the conditional probability that another person from the new city who avoids parties would fall in the category of shy people.

The transfer judgments are best described in terms of two experimental variables. First, the proportion of times that a description appeared in a category in the new city, in the training phase, was examined at five levels from 0% to 100%. For example, the proportion of people who avoided parties in the new city that had appeared in the *shy* category was varied

from 0% to 100%. Second, half the test questions involved a pairing that was congruent with prior knowledge, such as people who avoid parties being shy. Half the test questions involved a pairing that was incongruent with prior knowledge, such as persons smiling more than average falling in the category of people who are unhappy.

Heit's results (1994) clearly showed influences of prior knowledge as well as observed category members. For example, Fig. 2 shows the outcome of experiment 2, with the average responses indicated by points on the graph. The lines labeled as congruent refer to conditional probability judgments between features that are congruent with each other according to prior knowledge, such as a judgment of the likelihood that someone who avoids parties will be in the *shy* category. Likewise, the lines labeled as incongruent refer to probability judgments between features that are incongruent with each other according to prior knowledge, such as a judgment of the likelihood that someone who smiles a lot is in the category of unhappy people. The *X* axis indicates the observed proportion of category membership, in the training trials, corresponding to the test question. There was a clear main effect of prior knowledge as well as a clear effect of observed proportion, with higher observed proportions obtaining higher estimates.

Notably, there was no statistical interaction between the two experimental variables. This result suggests that people were simply summing up two sources of information: prior knowledge and observations. Heit (1994) implemented this idea with the integration model, a variant of exemplar models of categorization in which prior knowledge is represented by a number of prior examples. Subjects' prior knowledge of shy people might

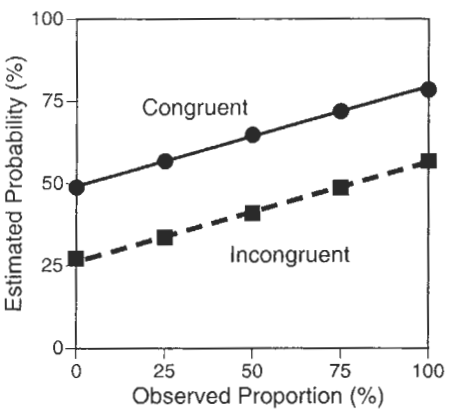

Fig. 2. Results of Heit (1994, Experiment 2). Reprinted by permission of APA.

be thought of in terms of a number of prior examples of shy people from outside of the experiment. Knowledge about shy people from outside of the experiment, such as that they avoid parties, will be transferred to the experimental context when a new group of shy people are observed. That is, the representation of the category of shy people in the new city reflects observations of actual category members as well as prior examples derived from observations of shy people in other contexts. The predictions of the integration model, shown as the lines in Fig. 2, were obtained by simply estimating a fixed number of prior stereotypical examples retrieved for each category, and including these prior examples in the application of the model. The gap between the congruent and incongruent lines in Fig. 2 reflects the influence of these prior examples.

Further work (Heit, 1995) investigated the time course of the integration process by testing subjects at various points during the course of learning (i.e., after increasing numbers of category members had been observed). Subjects were asked to make transfer judgments about people in city *W* after 0, 4, 8, 12, or 16 observations had been made per category. The results are shown in Fig. 3. What is notable is that at each point during learning, the judgments seem to be derived from a simple combination of prior knowledge and observations, with prior knowledge determining judgments completely when 0 observations had been made, and diminishing influences of prior knowledge with more observations. This point is evidenced by the decreasing distance between the congruent and incongruent lines in successive graphs. Heit (1995) explained these results in terms of subjects retrieving a fixed number of prior examples as a starting representation for each novel category, then updating the category representation with new observation are made. The lines represent the predictions of an exemplar model of categorization that embodies this explanation.

C. Background Knowledge Used to Select Observations

Although the integration process documented by Heit (1994, 1995) seems to be prevalent (see also, e.g., Hayes & Taplin, 1995; Ward, 1994), there are other ways that knowledge affects category learning. In fact, most past research on prior knowledge effects on category learning has probably emphasized other effects rather than integration. Some researchers (e.g., Keil, 1989; Murphy & Medin, 1985; Murphy & Wisniewski, 1989) have argued that *selective weighting* effects of prior knowledge are critical in category learning. That is, previous knowledge leads us to attend selectively to certain features or certain observations during concept learning, thereby simplifying the task.

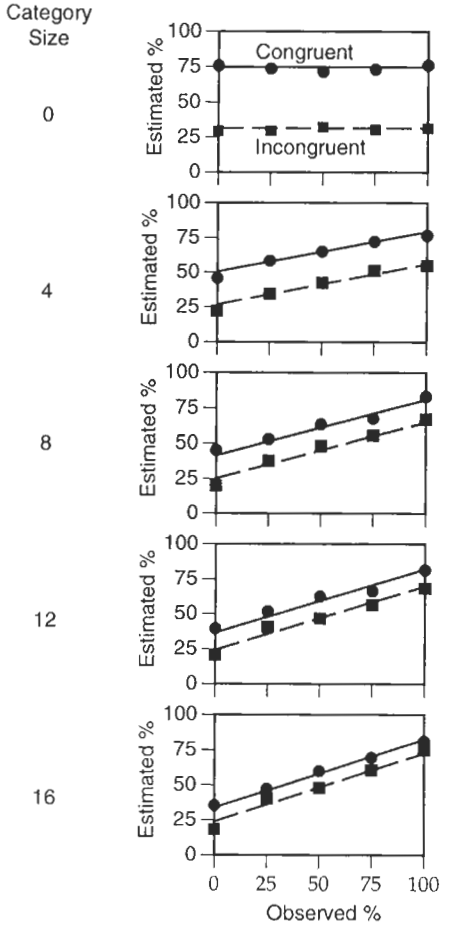

Fig. 3. Results from Heit (1995).

Many studies of category learning have obtained results that can be explained in terms of selective weighting (e.g., Keleman & Bloom, 1994; Medin, Wattenmaker, & Hampson, 1987; Murphy & Wisniewski, 1989; Wisniewski, 1995). For example, Medin et al. (1987) used a sorting task to study how people construct categories. Medin et al. found that when people sorted items into groups, they were especially likely to be influenced by pairs of dimensions that were causally related according to prior knowledge. For example, in sorting medical patients who were described by several symptoms, subjects were likely to sort on the basis of a pair of related

symptoms such as dizziness and earache, presumably because these dimensions were given extra weight.

More recently, Heit (1998a) found selective weighting of category members rather than features. On the basis of this research, it appears that not all category members are treated the same. Instead, some observations have greater influence on learning than others. Using a similar procedure (Heit, 1994, 1995), Heit (1998a) found that when training occurred at a slower pace, with more than 10 s per observation, there was an interaction between the effect of prior knowledge and the effect of observations, with the effect of prior knowledge being reduced for mixed observations (near the 50% range) compared to unmixed observations (near the 0% and 100% range). Figure 4 shows representative results from Heit (1998, experiment 1). The lines in the figure show predictions of a categorization model incorporating two distinct effects of prior knowledge. First, prior knowledge provides an initial set of expectations, as represented in the model by a set of prior examples. Second, category members that are incongruent with prior knowledge are selectively weighted and have a greater influence on categorization than theory-congruent category members. The curvature of the lines indicates that different observations, congruent or incongruent, are having different influences. The selective weighting in favor of incongruent category members tends to reduce the initial effect of prior knowledge, particularly with mixed observations (around 50% congruent). Hence the lines move closer together near the middle of the graph. This finding suggests that given enough time to process category members, people will apply strategic processes, allowing incongruent observations to have a greater influence.

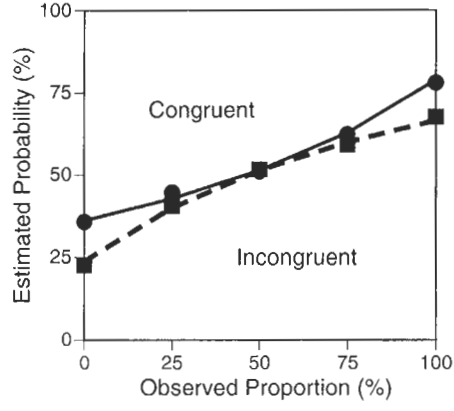

Fig. 4. Results of Heit (1998, Experiment 1). Reprinted by permission of APA.

Generally speaking, it now seems well-accepted that background knowledge has a selective role in category learning, in a number of ways, just as this point has been made in related areas of research such as memory (e.g., Alba, 1983) and reasoning (e.g., Wason, 1960).

D. Concepts Used as Input to Background Knowledge

Wisniewski and Medin (1994) have argued that knowledge-driven processing and data-driven processing must be tightly coupled. That is, information should flow in both directions. They demonstrated this point in a set of studies in which subjects' background beliefs about categories such as creative children had to be adapted when observing stimuli such as ambiguous drawings done by children. It appeared that the subjects used the stimuli to acquire more general knowledge about how to parse drawings into features. (See also Schyns, Goldstone, & Thibaut, 1998, for an extensive discussion of how people learn to represent categories in terms of features.) This point regarding the flow of information from newly learned concepts to background knowledge was demonstrated in another way by Heit (1994), who, following a standard procedure of teaching subjects about people in city *W*, asked the subjects to make background judgments about people in the whole state rather than just in this one city. Figure 5 shows sample results, adapted from Heit (1994, experiment 5). There were large effects of prior knowledge, which was not surprising given that subjects were asked a general knowledge question. However, the slope of the lines also indicates that what the subjects observed in city *W* (just eight observations per

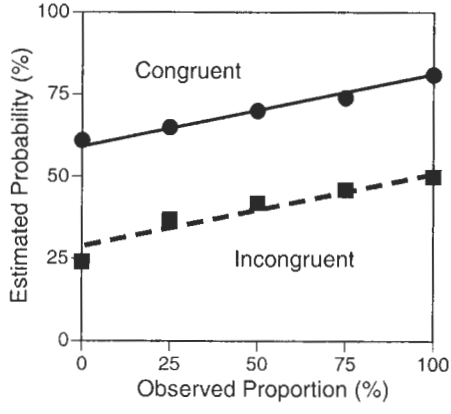

Fig. 5. Results of Heit (1994, Experiment 5). Reprinted by permission of APA.

category) had a substantial effect on their background knowledge judgments as well. The fact that subjects were tested immediately after they had observed the descriptions of people in city *W* could have led to an amplification of this effect. However, a longer delay between study and test may have made source discrimination even more difficult, as in a sleeper effect (e.g., Hovland & Weiss, 1952). Therefore, it is unclear to what extent Heit's procedure (1994) magnified the influence of new concepts on background knowledge.

Heit (1994) accounted for the effects of recent observations on background knowledge judgments in the same way as the effects of background knowledge on a recently learned concept. In both cases, the integration model was applied, making the assumption that a categorization judgment would depend on the retrieval of memories for observations, corresponding to background knowledge and members of a recently learned category. The only difference in the accounts of the two cases was that for background knowledge judgments, it was assumed that a greater proportion of the retrieved memories would correspond to background knowledge compared to the judgments about a new concept. Predictions of the integration model are shown as the lines in Fig. 5. (For another discussion of revision of background knowledge, emphasizing rule-based systems, see Mooney, 1993.)

E. Observations Used to Select Background Knowledge

Until now in this chapter, a crucial issue concerning knowledge effects on category learning has been passed over. At the beginning of the chapter, we argue that people face the problem of too many individual cases, so they treat individual things as belonging to categories. Yet this solution raises another problem—that there can be an extremely large number of ways to group a set of individuals into categories. This problem can be addressed, we argue, by using background knowledge to constrain category learning. Unfortunately, this solution itself raises yet another problem; namely, the problem of selecting prior knowledge. Just as there are many individual observations to deal with, and many possible category structures that could be considered, there are many possible sources of background knowledge that could be helpful in learning about a new category. For example, imagine visiting a new town or university campus and looking at the buildings there, trying to learn about the general layout and architectural styles. Many sources of background knowledge could possibly be helpful, such as memories of other towns or other campuses. In fact, it would be easy for the number of past observations to greatly outnumber the number

of new observations! Even if past observations are organized and summarized, into a smaller number of categories, there will still be information corresponding to many different places and many different kinds of buildings. How could a person select useful information from all of this background knowledge and, in light of this knowledge selection problem, how could background knowledge actually make concept learning easier?

On the surface, the knowledge selection problem would seem very troublesome for experimental and computational approaches to category learning and influences of past knowledge. It would be easy to justify not doing research on this topic. Although the knowledge selection problem does seem very imposing, and potentially unsolvable, it is still important to note that people do solve this problem every day. People face new situations and they manage to retrieve useful background knowledge somehow. In spite of the large numbers of things to observe, possible categories to put them in, and possible sources of background knowledge to guide this categorization, people are not normally left helpless due to issues in computational complexity. Therefore, we do see the knowledge selection problem as an appropriate issue for empirical study; namely, we are interested in how people find useful prior knowledge for category learning from the many possible sources of prior knowledge. In addition, it is encouraging to pick up any textbook on Bayesian statistics (e.g., Raiffa & Schlaifer, 1961) and find many techniques listed for combining multiple prior beliefs with observations, and selecting among these beliefs based on the data observed. In Bayesian statistics there is no assumption that a learner starts with optimal or perfectly correct prior beliefs. Instead, the learner begins with a reasonable guess that merely serves as an initial basis for learning, with corrective information then provided by the data. Indeed, it is possible to start with a whole set of different prior beliefs, with a distribution of initial degrees of confidence in each of these. When observations are made, confidence in various prior beliefs can be increased or decreased as appropriate (see also Heit, 1998b). That is, observations can be used to select from among a set of prior hypotheses. Therefore, Bayesian statistics already does provide an approach for addressing the knowledge selection problem, and indeed, our own categorization model to be proposed in this chapter takes some ideas from the Bayesian approach. Still, it might be argued that even Bayesian statistics does not fully address the knowledge selection problem because these methods merely indicate how to select among a set of prior hypotheses, but they do not say which prior hypotheses should be chosen. The key point is that Bayesian techniques can be applied to a large set of prior hypotheses, even when many of them are highly abstract, repetitive, or even ill chosen, as long as this set covers the hypothesis space well enough so that the target concept can be represented.

Many previous experiments on knowledge effects on category learning, including Heit (1994, 1995, 1998a), have avoided the knowledge selection problem by more or less telling the subjects which prior knowledge to use in learning new categories. For example, when subjects learned about shy people in city *W*, it was easily understood that they were supposed to use prior knowledge of shy people in the real world. In contrast, some experiments have given subjects a more difficult task, using unlabeled categories or nonsense labels that minimize the clues available that might indicate which prior knowledge might be useful (e.g., Murphy & Allopenna, 1994; Wisniewski, 1995).

For example, in Murphy and Allopenna (1994), subjects learned about categories of animals, vehicles, and buildings, with labels such as "Category 1" and "Category 2." These labels obviously did not constrain the knowledge selection problem very much. When a subject learned about a category of vehicles, for example, there were many known types of vehicles that could be informative. It was impossible to know in advance whether to use prior knowledge about snowmobiles, ice cream vans, heavy trucks, or jeeps. However, the content of the category itself—that is, the descriptions of category members—were helpful in finding useful prior knowledge. For example, when subjects observed a category member with the description "Made in Africa, lightly insulated, and drives in jungles," they were able to access knowledge about vehicles used in hot weather such as jeeps, rather than knowledge about other vehicles such as snowmobiles and heavy trucks. This process is denoted in Fig. 1 by the arrow running from observations to background knowledge. In these experiments, subjects had so much possible prior knowledge to apply to category learning that they needed to use the observations themselves to select and assemble helpful prior knowledge.

Our own experiments were an attempt to further address the phenomenon of knowledge selection for category learning. Like Murphy and Allopenna, we used building categories (in experiment 1) and vehicle categories (in experiment 2). Given the extensive range of background knowledge people have for these domains, and the many familiar categories within these domains, we see these stimuli as encouraging knowledge selection processes. Unlike Murphy and Allopenna, we collected data over the course of learning. It seemed valuable to look at knowledge selection processes as they unfold over time. One of our goals was to show that in some situations, categorization judgments are not affected early on by prior knowledge until many observations have been made and relevant prior knowledge can be assembled—the opposite result of Heit (1995). Therefore, it was necessary to collect categorization judgments after various numbers of category members had been observed. Another advantage of

collecting data along the course of learning was that our data were suitable for developing and testing a computational model of category learning. The greater number of data points compared to Murphy and Allopenna's experiments provided a more constraining data set for modeling.

Our general prediction for these experiments was that, in terms of various measures, there would be increasing knowledge effects over the course of learning because subjects would have no indication, at the start of learning, which of many sources of prior knowledge would be relevant. We see this as a useful area of empirical study because most past experiments in this area just have not addressed the time course of prior knowledge effects. More important, a major class of models would make just the opposite prediction—namely, that prior knowledge would have its greatest influences early on, and these influences would be reduced over the course of learning. This prediction is made by "knowledge-first" categorization models, such as the integration model of Heit (1994), that have an initial store of prior knowledge, represented as exemplars, rules, prototypes, or connection strengths, and simply revise this representation to reflect local conditions. Early on, prior knowledge dominates judgments because that is the only information available. However, error-correcting learning mechanisms would lead to a more veridical representation over time, diminishing any influences of prior knowledge.

We next present our two experiments on knowledge selection in category learning, followed by a more general review of computational models that employ prior knowledge and then by the introduction of a new computational model that addresses knowledge selection effects.

II. Experiment 1

A. Method

In this first experiment, the 77 subjects learned about two categories of buildings, referred to as "Doe buildings" and "Lee buildings." The subjects were told to imagine that they were reading a book with a series of descriptions, each corresponding to a different building. The stimuli were organized in five blocks, with descriptions of four Doe buildings and four Lee buildings presented in each block. Each description included the category label (Doe or Lee) and a list of featural information, presented in a randomized order. There were two critical features presented in each description and two filler features. The critical features for each category were related to a known type of building (e.g., churches for Doe and office blocks for Lee or vice versa). The filler features, arbitrarily assigned to each category, were general

characteristics that could be true of just about any building. Finally, each description contained three pieces of individuating information (name of builder, surveyor, and photographer). This information was included simply to make the descriptions a bit longer and more difficult so that learning did not occur too quickly. Results for the individuating features are not reported here.

The critical and filler features were derived from a pretest. The object of the pretest was to ensure that the critical features would be grouped together consistently to form two categories and that the filler features would be distributed evenly between these two categories. The pretest involved a series of sorting tasks in which subjects were asked to place each feature into one of two groups. (Subjects were not given category labels for the two groups; instead, they freely sorted cards with feature names into two piles.) Initially, there were 18 pairs of binary features: 9 intended to be critical features and 9 intended to be filler features. For successive runs of the pretest, critical features were dropped or replaced if subjects did not show a strong preference for putting them in one category, and likewise filler features were dropped if subjects did show a strong preference for one category or the other. After a series of iterations of this procedure, a set of 8 pairs of critical features and 8 pairs of filler features was obtained. A final pretest group of 20 subjects sorted each of the critical features with at least 90% preferring one group over the other, and for the filler features preference for one group was always less than 75%. In addition, subjects were readily able to describe one sorted pile of features as being related to churches or old buildings, and the other as being related to office buildings or other commercial buildings. The complete list of critical features as well as sample filler features is shown in Table II.

From the 8 pairs of critical features, 4 pairs were randomly assigned to presentation frequency one. Each feature in each pair was presented in one description per block, either Doe or Lee. Two pairs were assigned to presentation frequency two, and each feature presented in two descriptions per block. Finally, two pairs of features were not presented at all in the study blocks (but they were tested in test blocks).

The whole experiment was a sequence of five study–test blocks. In each study block, the building descriptions, each with a category label, were presented individually, for 6 s each. A sample description would be: {Lee building type, Builder: T Jones, near a river, has gas central heating, Surveyor: R Rawson, Photographer: A Ferraro, has steeply angled roof, has wooden furniture}. Subjects were instructed to try to memorize the stimuli. Following each study block was a test block in which subjects were asked to categorize 40 single features in the Doe or Lee categories. These test items included 24 individuating features, 8 critical features (4 presented

TABLE II

CRITICAL AND FILLER FEATURES FOR BUILDING STIMULI

Critical features
Has steeply angled roof
Has wooden furniture
Has an interesting structure
Old building
Quiet building
Lit by candles
Ornately decorated
Built with stone
Has a flat roof
Has metal furniture
Has a repetitive structure
New building
Busy building
Lit by fluorescent light
Blandly decorated
Built with metal and concrete
Sample filler features
Near a bus station
Designed by a local architect
Has gas central heating
Not near a bus station
Designed by an international architect
Has electric central heating

once, 2 presented twice, and 2 not presented), and 8 filler features (same distribution as critical features). Overall accuracy feedback was given at the end of each test block to encourage good performance.

B. RESULTS

Initial analyses did not reveal any significant differences between presentation frequency 1 and presentation frequency 2; therefore, the results were pooled over these two presentation frequencies. The average proportions correct are shown in Fig. 6. The top panel shows responses to features that had been presented during the study blocks. Overall, there is a trend for performance to improve over blocks. Although there is no difference between critical and filler features in the first block, the

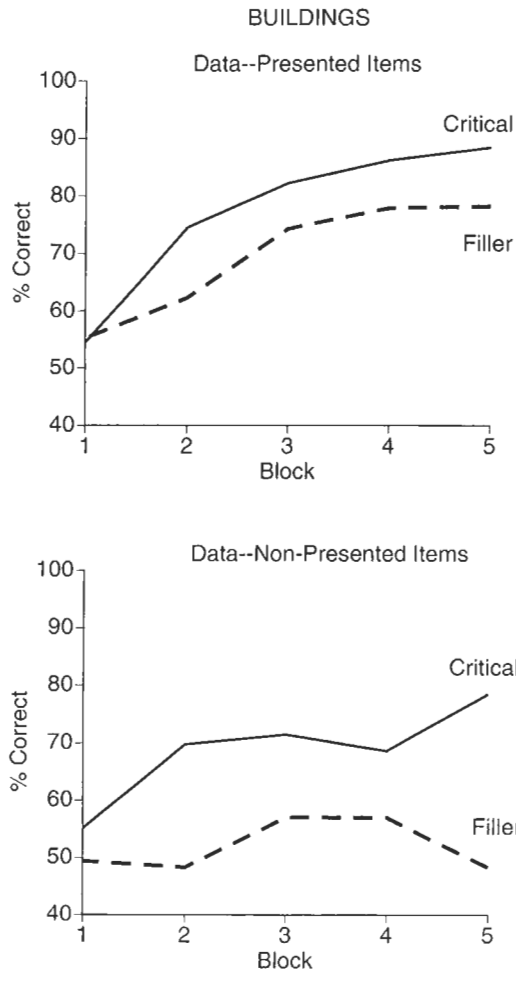

Fig. 6. Results for Experiment 1.

difference between the two kinds of features—that is, the gap between lines—widens after the first block. The bottom panel shows responses to the features that had not been presented at all. Here, observation of features cannot be playing a part in responses; any responses would be due to activation of prior knowledge. Responses to filler features essentially represent chance responding. The responses to critical, nonpresented features are more interesting. Even though these features were

never presented in study blocks, categorization performance clearly improved from the first block to the fifth block.

The results were analyzed with a three-way ANOVA with block, feature type (critical or filler), and presentation (observed or not observed) entered as variables. Each of the variables had statistically significant main effects, and likewise each of the two-way interactions were significant. Perhaps the most important interaction was the feature type by block interaction, supporting the observation that the difference between critical and filler features increased across blocks.

III. Experiment 2

A. Method

This experiment was intended to be a replication of the first experiment with a different stimulus set (vehicles rather than buildings). The main procedural change was that the experiment had six study–test blocks rather than five, in an effort to get a fuller picture of the course of learning. The critical and filler features were derived from a pretest in a similar manner to experiment 1. One set of critical features was intended to be related to tractors and the other was related to racing cars. The critical features as well as sample filler features are shown in Table III.

B. Results

Again, there was not any significant effect or interaction due to presentation frequency of features (once or twice per block), so the data were pooled over these two presentation frequencies. The results, in terms of average proportion correct, are shown in Fig. 7. Again, the pattern is for performance to improve with increased training, for people to be more accurate on critical features than filler features, and for the difference between critical features and filler features to increase over time. For example, on presented features there is a 10% difference between critical and filler features in block 1, but a 22% difference in block 4. The advantage of critical features over filler features is diminished somewhat by block 6, but this result may be due to a ceiling effect on critical features. Also, on the nonpresented features, there is steady improvement on critical features from block 1 to block 6 (and judgments on filler features again represent chance guessing). The results of a three-way ANOVA were similar to that of the first experiment, in that each of the three main effects (block, type of feature, and presentation) as well as the two-way interactions were statistically significant.

TABLE III

CRITICAL AND FILLER FEATURES FOR VEHICLE STIMULI

Critical features
Useful for pulling heavy objects
Is very heavy
Used for doing work
Drives on dirt roads
Uses diesel
Driver sits high off the ground
Not aerodynamic
Goes slowly
Not useful for pulling heavy objects
Is very light
Used for entertainment
Drives on smooth roads
Uses petrol
Driver sits close to the ground
Aerodynamic
Goes fast
Sample filler features
Has a rectangular gearbox
Tires made of synthetic rubber
Has gas shock absorbers
Has a spherical gearbox
Tires made of natural rubber
Has hydraulic shock absorbers

IV. Discussion of Experiments

The similarities between these two experiments are more important than the differences. In both experiments, subjects were increasingly influenced by background knowledge over the course of learning, in contrast to the results of Heit (1995). One source of evidence for increasing influences of knowledge is the results for presented features, in the top panels of Figs. 6 and 7. For the building stimuli, there was no difference in classification accuracy for critical and filler features after the first training block, but by the end of the second block subjects had apparently retrieved prior knowledge that facilitated performance on critical features compared to filler features. Realizing that the Doe buildings are churchlike and the Lee buildings are like office buildings, for example, would help answer questions

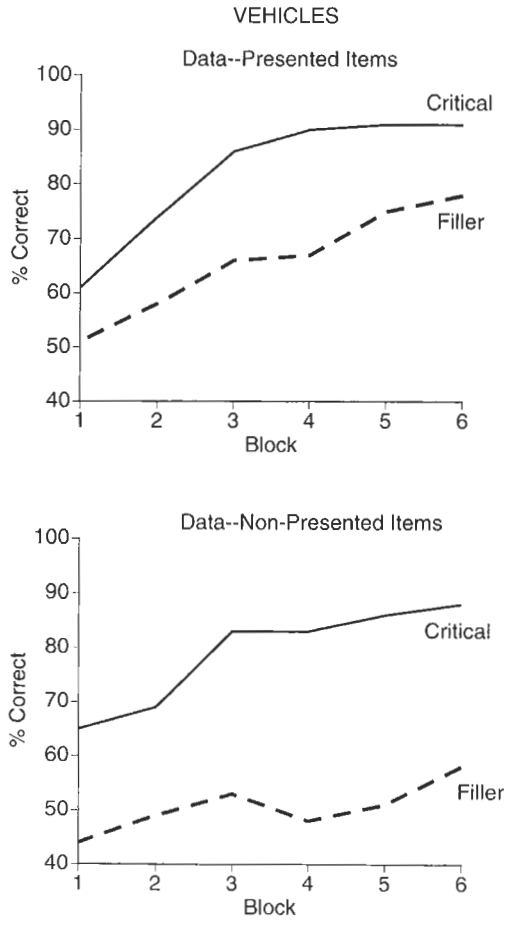

Fig. 7. Results for Experiment 2.

about critical features but not filler features. Although performance on critical and filler features continued to improve over the course of learning, the advantage for critical features was persistent. The results for vehicles were similar, except that there was an advantage for critical features even after the first block. Perhaps for these stimuli, seeing just four observations per category was enough to retrieve some useful prior knowledge. It is possible that if we had tested subjects halfway through the first study block of experiment 2, the results would have been more similar to experiment 1. In addition, for the vehicle stimuli, the advantage for critical features over filler features increased over time, more than doubling from the first block to the fourth block.

The other source of evidence for changes in knowledge effects is the judgments on nonpresented critical features, shown in the bottom panels of Figs. 6 and 7. Subjects were never told the correct category for these features during training blocks. The only way to classify these features correctly was on the basis of general knowledge (about buildings or vehicles). In both experiments, performance on nonpresented critical features improved over the course of learning, suggesting that subjects were increasingly relying on appropriate knowledge for making judgments about these features.

Why were the results of these experiments so different from those of Heit (1995)? Why do prior knowledge effects sometimes increase with learning and other times decrease with learning? The main difference between the present experiments and Heit (1995) is that in the present experiments, the category label names (e.g., Doe building type) did not suggest any particular source of prior knowledge, whereas in Heit (1995), the categories (e.g., shy people in city *W*) readily suggested which prior knowledge should be used. The Heit (1995) experiments failed to detect any increased use of prior knowledge over learning because there was an initial ceiling effect—the relevant prior knowledge was so easily retrieved at the start of the experiment, there was no chance for its influence to increase any further. Why didn't the present experiments find less use of prior knowledge over time? Indeed, there was a persistent advantage for critical features over filler features, even in blocks 5 and 6. It is hard to say whether performance on presented filler features would ever come up to the level of presented critical features, even with much more training. It seems likely that continued testing of individual features interleaved with training blocks would encourage subjects to learn about as many features as possible, but practical matters such as greater levels of motivation in early blocks compared to later blocks might make it difficult for filler features to ever be learned as well as critical features.

One surprising result, or lack of result, from these experiments was the lack of difference between features presented once per block and features presented twice per block. For both critical and filler features, we did not find any statistically significant difference in judgments for the two levels of presentation, despite the 100% difference in presentation frequency. It is tempting to relate this finding to results from Murphy and Allopenna (1994), who also found low sensitivity to frequency manipulations for stimuli that lead to retrieval of prior knowledge. However, it would be wrong to conclude that people are not sensitive to frequency information when category learning involves prior knowledge. For example, Heit (1994, 1995, 1998a) documented a very robust pattern of responses to variations in frequency of presentation (see Figs. 2–4). Also, informal debriefing of

subjects suggested to us that because each description, containing eight pieces of information, only appeared for 6 s, there may have been some strategic scanning of information. For example, in each block some subjects might have looked for features that had not already been presented in that block to maximize the amount of fresh learning per block. So the effect of a second presentation of a feature within a block could have been diminished due to some subjects' learning strategies. Therefore, we find the lack of frequency effects interesting, but it seems to require further study before stronger conclusions are reached. Indeed, Spalding and Murphy (in press) have argued that the lack of sensitivity to frequency in Murphy and Allopenna would depend on the judgment task being used (e.g., classification or frequency judgment).

V. Putting Knowledge into Neural Network Models

Having collected some data on the time course of knowledge selection in category learning, we set out to develop and apply a computational model that could address these phenomena. Previous modeling efforts (Heit, 1994, 1995, 1998a) did not address knowledge selection at all. Rather than continuing along these lines of extending the framework of exemplar models, we decided to develop a new model within the framework of connectionist or neural network models. Although exemplar models have some advantages, such as their simplicity and their wide success in application to categorization data, connectionist models seem to provide a richer descriptive framework. That is, the greater complexity of connectionist models in terms of possibilities for different architectures, learning rates, activation rules, initial connection weights, and so on, provides more opportunities for describing distinctive effects of knowledge on learning, as well as an appropriate framework for describing the dynamics of learning and the interplay of knowledge, concepts, and observations. Also, there has already been a great deal of research, mainly outside of psychology, on different ways of putting knowledge into neural networks. Before we present our own model, we review some of this past work, largely from the field of engineering.

A useful framework for discussing prior knowledge in neural networks has been developed by Geman, Bienenstock, and Dourstat (1992). In their discussion of computational models of learning, they demonstrated that the generalization error when learning a concept can be broken down into a *bias* component and a *variance* component. Models that rely heavily on prior assumptions about the data (e.g., having architectural constraints that favor a particular conceptual structure) can lead to a high bias component;

that is, the model can persistently fail to capture aspects of the target concept that do not meet its prior assumptions. However, models that do not make strong assumptions about the concept to be learned can show a high variance component; that is, that they will be easily swayed by noise in training samples. Therefore a model without many assumptions could require an excessively large training sample to achieve satisfactory generalization performance. Furthermore, reducing one type of error frequently is accompanied by an increase in the other type of error, leading to what Geman et al. (1992) referred to as the *bias–variance dilemma.* To reduce generalization error, both bias and variance must be reduced. One way of doing so would be to increase the number of training examples. Unfortunately, as Geman et al. show, in practice the number of training examples will be insufficient to achieve anywhere near optimal performance. We next review a number of learning algorithms from artificial intelligence (AI) research that are aimed at reducing generalization error, keeping in mind the need to minimize the number of training examples as well.

One method for reducing the number of examples required for good generalization is to introduce "hints" into neural networks (Abu-Mostafa, 1993, 1995). Hints are general properties of a class of target concepts, independent of the specific details of the training data. For example, a hint in letter recognition might be that the mapping of a pixel image of a letter to the identification of that letter is position invariant. Hints are introduced into the network by presenting "virtual examples" of the hint and altering the error function to incorporate a term for the hint. (There is some similarity between virtual examples and "prior examples" in Heit, 1994.) Building on the work of Vapnik and Chervonenkis (1971), Abu-Mustafa has derived a theoretical framework for predicting how much a particular hint will reduce the need for training examples.

Another approach to prior knowledge is to insert biases directly into neural networks by setting the weights before learning begins. This approach has been taken by, for example, Frasconi, Gori, and Soda (1995) and Giles and Omlin (1993). In both cases the specific method was to insert transition rules into recurrent neural networks; known transitions were built into the network and then unknown transitions were learned from the data. Giles and Omlin showed that "malicious" rules or incorrect prior knowledge could be overcome gradually by corrective training data. As Frasconi et al. (1995) noted, however, a potential problem with this method is that the longer a network is trained, the more likely it is to use a solution based on the data, thereby forgetting its prior knowledge. Frasconi et al. suggested a compromise of allowing the weights to vary within a constrained space, which was the technique employed by Choi, McDaniel, and

Busemeyer (1993). Also, rather than inserting knowledge directly, it is possible to train the network in one input–output domain and then rely on this prior knowledge to help learning about structurally similar domain, freezing a subset of the hidden units to prevent forgetting (Dienes, Altman, & Gao, in press).

We next review ways of building in prior knowledge by varying the network architecture. The basic goal here is to allow the network to have sufficient representational power to capture the underlying concept, but at the same time to avoid fitting the noise in the data. This goal is another way of looking at the bias–variance dilemma—a network that is too small leads to a high bias, but a network that is too large leads to high variance (and fitting the noise). Constructive networks (e.g., Giles, Chen, Sun, Chen, Lee, & Goudreau, 1995; Mareschal & Schultz, 1996; Prechelt, 1997) expand their architecture during learning, allowing the complexity of the network to increase as the data suggest it. Destructive networks, however, start off with an excess of hidden units and then prune off the hidden units that are not useful (e.g., Mozer & Smolensky 1989; Reed, 1993; or, for a more biological treatment, Brown, Hulme, Hyland, & Mitchell, 1994). The advantage constructive networks have is that they might require less computation than destructive nets and that there is no need to make an initial guess at the appropriate number of hidden units (Giles et al., 1995).

Rather than varying the network architecture over the course of learning, another approach is to employ more than one architecture within a mixed network and allow the network itself to learn which of the architectures is best for a particular problem. An example of this approach is the mixture-of-experts network (Jacobs, 1997; Jacobs, Jordan, & Barto, 1991; see also Erickson & Kruschke, 1998). For example, Jacobs et al. (1991) used a mixed network, with three modules having different structures (no hidden units, medium number of hidden units, and a high number of hidden units). In effect, each module took a different approach to the bias–variance dilemma, with the simplest network being most constrained in terms of what it could learn and the network with many hidden units being most sensitive to variation in a training sample. The network was trained to perform two tasks: an object localization task and an object recognition task. The localization task was simpler in that it did not require hidden units for good performance. The mixture-of-experts network learned to allocate the module without hidden units to the localization task while it allocated one of the modules with hidden units to the recognition task. We see the mixture-of-experts approach as coming close to the Bayesian idea of starting with multiple hypotheses then selecting among them based on the data (and see Jacobs, 1995, for a more substantial comparison).

VI. The Baywatch Model

A. Overview

Our own approach to the knowledge selection problem has some parallels to the mixture-of-experts architecture, but instead of using modules with different structures, we used modules with different pools of pretrained knowledge. Therefore, our method also has some relations to techniques that insert prior knowledge directly into networks. Our own model, illustrated in Fig. 8, can be described as having one module or set of weights for strictly empirical learning. These weights do not get any pretraining. Then the model also has a set of experts that are pretrained to recognize different known categories. For example, a network for learning about buildings might have experts that can recognize different kinds of buildings such as churches, office blocks, restaurants, and schools. (Only two of these expert modules are illustrated in Fig. 8.) We refer to this model as the *Baywatch* model because it combines a general Bayesian approach to selecting among multiple sources of prior knowledge with an empirical learning component.

The Baywatch model is a feedforward network in which the input units represent the individual features and the output units represent the Doe and Lee category nodes. The two hidden units correspond to two expert

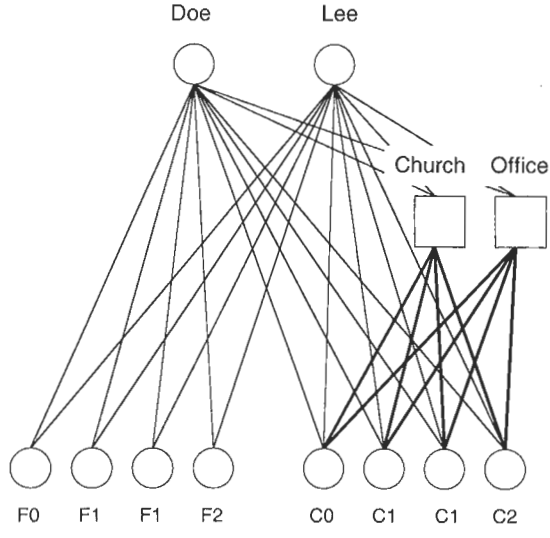

Fig. 8. Illustration of Baywatch model.

modules, or prior knowledge category nodes (PK nodes). The four input units on the left of Fig. 8 represent filler features, and the four inputs on the right represent the critical features. The only difference between the two types of features is that the filler features are only connected to the output nodes, whereas the critical features are connected both directly the output nodes and indirectly to the output nodes via the prior knowledge nodes. The difference between filler and critical features in the model reflects our assumptions about how learning would take place in our experiments. Consequently, we required filler features to be learned directly without the help of prior knowledge, whereas critical features were to be learned both directly and by a mediated connection through prior knowledge. The connections between the critical features and the PK nodes have fixed weights, so that values of critical features of the stimuli that correspond to church features would activate the church PK node, and likewise critical features of the stimuli that correspond to offices would activate the office PK node. It is assumed that these fixed weights would correspond to prior knowledge about familiar characteristics of churches and office blocks learned through ordinary means of association. The PK nodes have threshold functions, so that if any church feature, say, a steeply angled roof, is presented, then the church PK node will be activated. The activation from the PK node would then be propagated to the output units.

In contrast to the connection weights between the critical features and the PK nodes, the other weights in the network are learnable through gradient descent on the error between the desired output of the network and the actual output. Adjusting the weights from filler units and the critical units to the output units allows the features to be associated with the category nodes in the empirical learning module. Note that if these were the only weights in the network, there would be no difference between the two types of features. Finally, there are adjustable weights between the PK nodes and the category nodes. These represent the subject's capacity to associate known categories—say, churches and office blocks—with the new categories, Doe and Lee buildings. We see this part of the network as addressing (at least in part) the knowledge selection problem, because here the network is learning to select from already known categories and apply this knowledge to judgments about new categories. Finally, we note that same simulations were used to address experiments 1 and 2, which had the same stimulus structure and similar results. (We continue to refer to buildings rather than buildings and vehicles, for simplicity.)

B. Technical Details of the Model

The input units can take on the values $\{+1, 0, -1\}$, which correspond to the Doe value of a feature, the feature not being present, and the Lee

value of a feature respectively. For instance, if the feature is the lighting feature (see Table II), then a −1 value would mean "lit by candles" value, a 0 would correspond to not presenting the feature at all, and a +1 would mean "lit by fluorescent lights." The two output units vary continuously between −1 and +1. One output unit corresponds to the Doe category and the other to the Lee category. The activation on each category was given by the weighted sum of its inputs. This activation was then converted into a probability measure using the logistic transformation given in Gluck and Bower (1988, equation 7). If a Doe exemplar is presented during training, the teaching values for the category nodes are +1 on the Doe node and −1 on the Lee node (Table IV). These values would be reversed for a Lee training example.

Critical features are connected by fixed weights to the PK nodes. As can be seen from Fig. 8, these were connected so that if the Lee value (−1) of a feature is presented, this lead to positive activation on the church PK node (because Lee buildings would correspond to churches) and a negative activation on the office node. The output of a PK node was a threshold transformation of the weighted sum of its inputs, such that the output was 1 if the sum was greater than or equal to 1, and 0 otherwise. All of the weights in the network were adjusted according to the standard delta rule (e.g., Gluck & Bower, 1988).

C. Simulation of Experiments

The network was trained for a total of 10 epochs, with the learning rate in the delta rule set at 0.1 and the probability mapping constant for the logistic transformation function set at 7.0 (both values were derived from an informal sampling of the parameter space). The training stimuli consisted of four examples of buildings—two Doe exemplars and two Lee exemplars—which are shown in Table IV. The first two rows are the Doe buildings and the second two rows are the Lee buildings. Note that the fourth features in the critical feature section and in the filler feature section

TABLE IV

Structure of the Training Data

Filler features				Critical features				Desired output	
1	1	0	0	1	1	0	0	1	−1
1	0	1	0	1	0	1	0	1	−1
−1	−1	0	0	−1	−1	0	0	−1	1
−1	0	−1	0	−1	0	−1	0	−1	1

always have a value of zero. These features correspond to those that were never presented to the subjects in the experiments.

Following each training epoch, the network was tested on the individual features by presenting a vector of all zeroes except for the particular feature of interest, which had a value of either +1 or −1. The results of the simulations are displayed in Fig. 9, with the proportion correct on the test set shown as a function of the number of learning epochs and feature type. The top panel shows the model's predictions for presented features. As for the results of the experiments, the predictions for features presented once per epoch and features presented twice per epoch are pooled together.

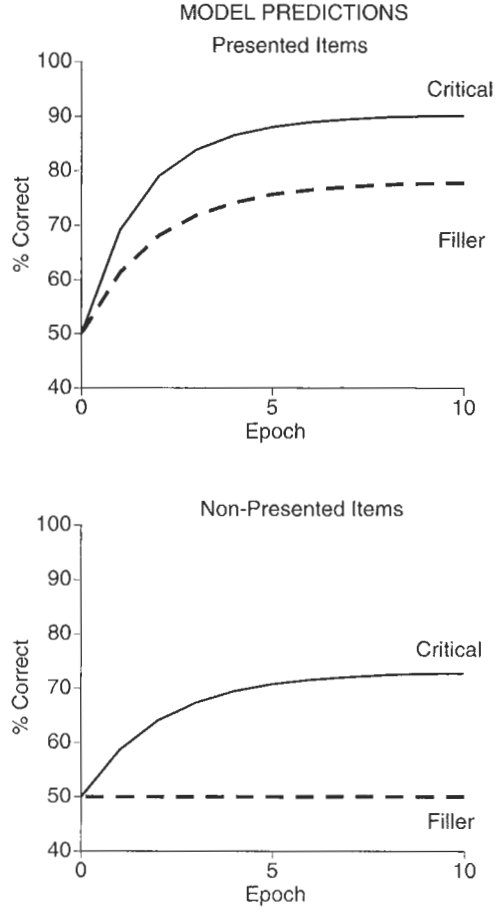

Fig. 9. Predictions for both experiments.

The bottom panel shows predictions for features that had not been presented during training. The predictions fit well with the main results of the experiments. Critical features were learned more quickly than filler features, and critical features that had not been presented were responded to more accurately than chance, whereas filler features that had not been presented were at chance level.

The first result can be explained in terms of the extra connections from critical feature inputs to the output units, mediated by connections through the PK nodes. As the network progressively learned which sources of prior knowledge correspond to the Doe and Lee categories, responses on critical features were derived both from the empirical learning module and from prior knowledge. In addition to these two paths of influence on the category outputs, the other advantage for critical features over filler features is that there are two paths of learning, in effect leading to twice as much updating of weights after a particular learning trial.

A similar advantage for presented critical features over presented filler features might be obtained without any PK nodes at all by simply increasing the learning rate on the critical features relative to the filler features. However that scheme would not predict any advantage for nonpresented critical features over nonpresented filler features. In the Baywatch model, for nonpresented critical features and filler features, the weights leading from the input units directly to the output units remain at zero throughout learning. Because this is the only way the filler features can activate the output units, their accuracy stays at chance level. In contrast, the nonpresented critical features have another route to the category units, through the PK nodes whose weights are adjusted when any critical feature are presented. Therefore the PK nodes are critical to the Baywatch model's predictions on nonpresented critical features.

To provide a better idea of how the Baywatch model uses prior knowledge, we reran the simulations without any PK nodes for comparison. In Fig. 10, we show simulated predictions on presented items, comparing versions of the model with and without prior knowledge. For critical features, in the top panel, it can be seen directly the prior knowledge does not have any influence initially on judgments; the model acts the same way with or without PK nodes. However, the beneficial effect of prior knowledge for critical features increases over the course of learning, as the network with PK nodes learns which categories to connect with its prior knowledge. In the bottom panel of Fig. 10, there is evidence for a slight detrimental effect of prior knowledge on the learning of filler features. This result can be explained as a kind of overshadowing effect, in which knowledge of some highly predictive cues can reduce learning on other predictive cues. As a consequence of the delta rule, when the network learns to predict the outputs increasingly well from the critical feature inputs, learning on

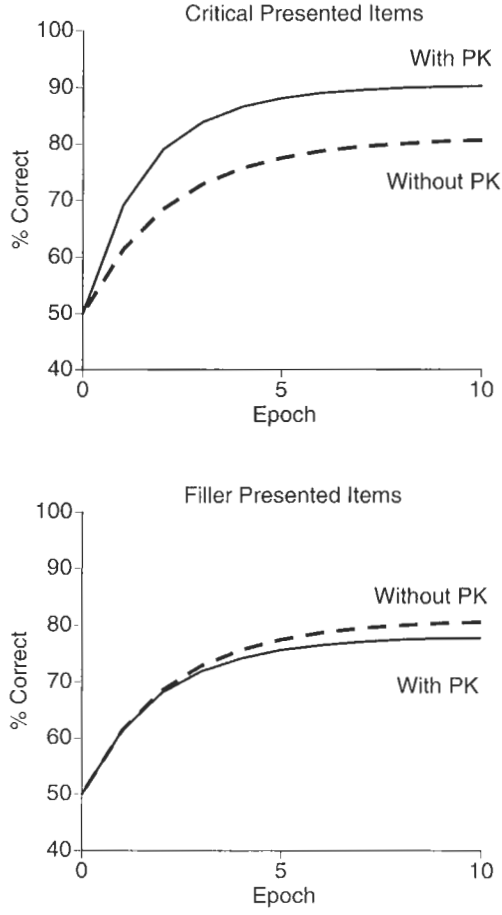

Fig. 10. Predictions with and without prior knowledge nodes.

the filler features will increasingly be disadvantaged. However, one possible difference between our experiments and the model is that the repeated testing of individual features could encourage subjects in the experiments to learn as much as possible about each individual feature regardless of how much is known about other features.

VII. Evaluation of the Baywatch Model

The Baywatch model captures many of the important features of the two experiments on knowledge selection in category learning. At the start of

learning, the model is not influenced by prior knowledge, because it does not know which past categories are useful for making predictions about the Doe and Lee categories. However, as observations are made, the model is able to select relevant prior knowledge to be used for judgments about the novel categories. This influence of prior knowledge leads to a persistent advantage for critical features over filler features. Admittedly, the Baywatch model would require more experimental testing before a complete evaluation can be made, but even this initial application brings up some interesting issues.

One notable difference between the model's predictions and subjects' performance is that the model would predict a robust effect of presentation frequency; that is, more accurate judgments for features presented twice per block compared to features presented once per block. (This prediction is not shown in Fig. 8, however.) In contrast, there was no significant difference between these two levels of presentation in the experiments. This insensitivity to frequency could be an important aspect of concept learning in knowledge-rich domains (c.f., Murphy & Allopenna, 1994), in which case it would be important to try to capture it in a future version of the Baywatch model. However, in the present experiments the lack of sensitivity to presentation frequency could just reflect subjects' reading strategies and might be highly dependent on number of features per presentation and the reading time allowed for each presentation. Therefore, further experimental study is required.

Perhaps a more fundamental question is to what extent the Baywatch model is really addressing the knowledge selection problem. The simulations were run with just two sources of prior knowledge (e.g., churches and office blocks) and the network was able to link up these two sources with the correct output categories, Doe and Lee. However, people would obviously have a much larger number of known categories when facing the knowledge selection problem due to large numbers of known kinds of buildings, vehicles, and so on. How well would the Baywatch model scale up? We think the model might scale up well, specifically in terms of adding more prior knowledge nodes. Our investigations so far have distinguished three different classes of PK nodes that might be added to the network in Fig. 8, in addition to the church and office nodes.

First, completely irrelevant prior knowledge nodes might be added that have little or no connection to the input stimuli. For example, there could be prior knowledge nodes for space stations, igloos, tents, and cave dwellings added to the network, but these nodes would hardly be activated by the inputs. For example, an input feature such as "lit by fluorescent light" would not be strongly associated with these categories, according to prior

knowledge. Therefore, adding PK nodes that are irrelevant to the stimuli would not affect the results of the simulations very much.

Second, additional PK nodes that are similar to the existing PK nodes might be added. For example, a PK node corresponding to cathedrals would entail much of the same connections to inputs as the church node. Likewise, there might be similar PK nodes for industrial parks and office buildings. In further simulations, we added a cathedral PK node that had two connections to the critical features for churches (to the critical feature presented twice and the nonpresented critical feature) and an industrial park PK node that likewise was connected to two critical features for office buildings. The results are shown in Fig. 11, comparing the original simulations with two PK nodes to the new simulations with four PK nodes. Inserting the two additional PK nodes improved performance on those critical features that now had two paths for knowledge-directed learning. However, inserting PK nodes did worsen performance on filler features because the additional reliance on critical features led to some overshadowing of filler features. Likewise, there was a slight decrement on performance (not shown in Fig. 11) on critical features that differed within a pair of PK nodes (e.g., features that were true of office buildings but not industrial parks). Still, to the extent that sources of prior knowledge were mutually supporting, having multiple sources of prior knowledge helped performance. Generally speaking, we did not find that adding additional similar PK nodes led to a knowledge selection problem. This result raises an interesting question about our experiments. Although we observed better performance on critical features than filler features, due to increased use of prior knowledge, the results themselves do not indicate *which* prior knowledge was being retrieved. Some subjects could well have been retrieving knowledge about cathedrals rather than churches, or industrial parks rather than office buildings. Indeed, informal debriefings of subjects revealed some variety of responses to questions about what the experimental stimuli were like in the real world.

Third, "malicious" prior knowledge nodes could be added to the network, for example, prior knowledge about some kind of building that is half-church and half-office block. Although we initially expected that malicious PK nodes would hurt performance, we had some trouble finding any negative effects in simulations. A half-church, half-office PK node would not get activated very much by our training stimuli, which after all did not contain any items that were half-church, half-office. To the extent that the malicious PK node did get activated, the network would learn equal associations between it and both the Doe and Lee output units. In sum, the malicious PK node was poor competition for real PK nodes, because it did not match the inputs well and it did not become strongly associated

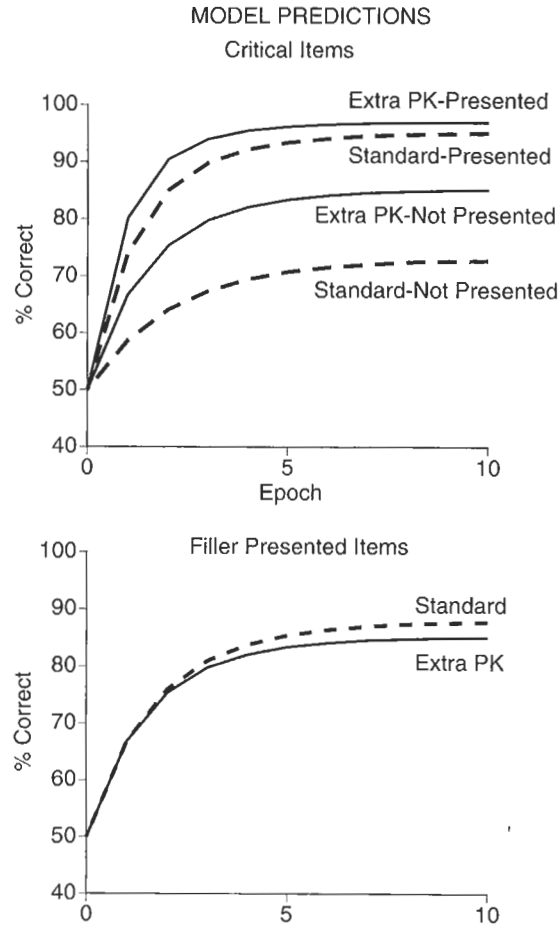

Fig. 11. Predictions with additional prior knowledge nodes compared to standard network.

with one output rather than the other. Again, we failed to find any knowledge selection problem due to adding malicious PK nodes. Of course, we intend to conduct further simulations involving additional PK nodes, but so far prospects look fairly good for the model's potential to be scaled up with more PK nodes and perform knowledge selection. The success of the Baywatch model in dealing with multiple PK nodes bears a great deal of resemblance to the ability of Bayesian statistics to work with multiple prior hypotheses, including some that are irrelevant, some that are repetitive, and some that are incorrect.

More generally, we see the knowledge selection problem as surely having many facets. Certainly one of them is that when learning about novel categories, a learner would need to link up knowledge of familiar categories with judgments about the novel categories. The Baywatch model seems to address this aspect of knowledge selection, in terms the gradual selection of prior knowledge nodes to use for a particular novel output category. In contrast, the prior knowledge in terms of connections from input units to PK nodes is fixed at the start of the simulations. It is assumed that these connections would have been already learned through ordinary associative processes so that the network can more or less instantly recognize church or office buildings. However, there could be some gradual aspects of knowledge activation or retrieval that are not captured by the model. It could be the case that somehow the connections between input units and PK nodes would be strengthened over the course of making observations so that the recognition of relevant categories in prior knowledge would not be instantaneous when a single observation is made. It could be valuable to study this aspect of knowledge selection more directly, for example by showing subjects a series of training examples and asking them to judge directly which familiar categories are related to these stimuli.

Finally, we would point out that the Baywatch model as presented in this chapter is but one possible variant within a larger class of models that could perform knowledge selection. For example, referring to Fig. 8, the model could have category label units (Doe and Lee) added to the input layer as well as feature units (F0, F1, etc.) added to the output layer, turning the model into an auto-associator. Such a model could make a greater variety of inferences, such as feature-to-feature inferences (e.g., Heit, 1992) in addition to the feature-to-category inferences in the present version of the model. Hence, the auto-associator version could be applied to a wider range of experimental tasks.

There are several other ways that the architecture of the Baywatch model could be modified. These changes were not necessary for fitting the results of our experiments, but they could be useful for application to other experimental designs. First, hidden units could be added to the empirical side of the network, allowing it to solve nonlinear classification problems. Second, the various modules in the network, including the empirical module and all the PK nodes, could be placed in greater competition with each other. The present architecture of Baywatch encourages cooperation between different modules, in the sense that outputs from multiple modules are combined to make a prediction. Instead, the network could be encouraged to specialize; for example, learning that different modules should be used for different stimuli. Some stimuli might be best classified with the empirical module alone, whereas other stimuli would be best classified based on a

single PK node. This scheme would force the network, for example, to choose between a church PK node and a cathedral PK node, rather than allowing their influences to combine. (See Jacobs, Jordan, Nowlan, & Hinton, 1991, for a further discussion of ways to increase competition between modules.)

Perhaps an even more radical change would be to alter the nature of the knowledge-driven side of the network. The knowledge-driven part of the network and all the PK nodes could be replaced by a module that has been pretrained with a set of rules for identifying buildings. This kind of architecture would make Baywatch closer to hybrid rule-plus-association networks such as those by Ashby, Alfonso-Reese, Turken, and Waldron (1998) and Erickson and Kruschke (1998). However, it is unclear to what extent such a network would make different predictions. Another, less extreme change to Baywatch would be to allow learning on the connections between critical input features and the PK nodes (again, see Fig. 8). At present, these connections are fixed at the start of learning, but it is possible that allowing these weights to change slowly would allow the network to address the issue of how global theories might change over time. That is, people may have a set of prior concepts that help learning, but these concepts themselves could be modified occasionally. To give a real example, one of the authors visited a church in Hungary that was in the shape of an owl; seeing this church led to learning about the local conditions as well as altering the author's general conception of churches.

A last extension to the Baywatch model, following Abu-Mostafa (1993, 1995), would be to apply it to situations in which the learner is given a hint about how to solve a classification problem. For example, a rather specific hint would be that Lee buildings are office buildings; such a hint could be given to the network in terms of pre-training and likewise this hint could be given to subjects in an experiment. The use of hints could be a good way to generate and test more detailed predictions of the Baywatch model. The model could be used to predict a hierarchy of hints, with some hints aiding learning more than others.

VIII. Conclusions

Since the influential Murphy and Medin (1985) paper that raised the issue of background knowledge in terms of category learning and models of categorization, there has been much progress on this issue (again, see Heit, 1997, and Murphy, 1993, for reviews). In particular, there has been a great deal of documentation of the various ways that prior knowledge influences category learning, for which Fig. 1 is only a partial summary. At present,

we see the most pressing and more exciting issue in this area of research to be the knowledge selection problem. On the surface it is a very discouraging problem, as it requires choices from many potentially useful sources of prior knowledge. It is easy to see why little research on categorization, from either experimental or modeling approaches, has addressed the knowledge selection issue. Yet people manage to solve this problem every day and use their prior knowledge profitably. Therefore we think it is important to address this problem head on, rather than avoiding it any longer. Our own approaches, involving experimental research on the time course of category learning and computational modeling of knowledge selection processes, are in their earliest stages but we are hopeful that these approaches will continue to be informative about this most important issue.

Acknowledgments

We thank Ulrike Hahn, Gregory Murphy, and Yves Rosseel for comments on this paper. This research was supported by the Economic and Social Research Council and the Biotechnology and Biological Sciences Research Council (United Kingdom) and the National Institute of Mental Health and National Science Foundation (United States).

Please address correspondence to Evan Heit, Department of Psychology, University of Warwick, Coventry, United Kingdom; email: E.Heit@warwick.ac.uk.

References

Abu-Mostafa, Y. S. (1993). Hints and the VC dimension. *Neural Computation,* 278–288.

Abu-Mostafa, Y. S. (1995). Hints. *Neural Computation, 7,* 639–671.

Alba, J. W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin, 93,* 203–231.

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

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

Brown, G. D. A., Hulme, C., Hyland, P. D., & Mitchell, I. J. (1994). Cell suicide in the developing nervous system: A functional neural network model. *Cognitive Brain Research, 2,* 71–75.

Choi, S., McDaniel, M. A., & Busemeyer, J. R. (1993). Incorporating prior biases in network models of conceptual learning. *Memory & Cognition, 21,* 413–423.

Dienes, Z., Altman, G., & Gao, S.-J. (in press). Mapping across domains without feedback. *Cognitive Science.*

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127,* 107–140.

Frasconi, P., Gori, M., & Soda, G. (1995). Recurrent neural networks and prior knowledge for sequence processing: A constrained nondeterministic approach. *Knowledge-Based Systems, 8,* 313–328.

Geman, S., Bienenstock, E., & Dourstat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation, 4,* 1–58.

Giles, C. L., & Omlin, C. W. (1993). Extraction, insertion and refinement of symbolic rules in dynamically driven recurrent neural networks. *Connection Science, 5,* 307–337.

Giles, C. L., Chen, D., Sun, G., Chen, H., Lee, Y., & Goudreau, M. W. (1995). Constructive learning of recurrent neural networks: Limitations of recurrent cascade correlation and a simple solution. *IEEE Transactions on Neural Networks, 6,* 829–836.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 227–247.

Hayes, B. K., & Taplin, J. E. (1995). Similarity-based and knowledge-based process in category learning. *European Journal of Cognitive Psychology, 7,* 383–410.

Heit, E. (1992). Categorization using chains of examples. *Cognitive Psychology, 24,* 341–380.

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

Heit, E. (1995). Belief revision in models of category learning. In *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 176–181). Hillsdale, NJ: Erlbaum.

Heit, E. (1997). Knowledge and concept learning. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts, and categories* (pp. 7–41). London: Psychology Press.

Heit, E. (1998a). Influences of prior knowledge on selective weighting of category members. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 712–731.

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

Hovland, C. I., & Weiss, W. (1952). The influence of source credibility in communication effectiveness. *Public Opinion Quarterly 15,* 635–650.

Jacobs, R. A. (1995). Methods for combining experts' probability assessments. *Neural Computation, 7,* 867–888.

Jacobs, R. A. (1997). Nature, nurture, and the development of functional specializations: A computation approach. *Psychonomic Bulletin & Review, 4,* 299–309.

Jacobs, R. A., Jordan, M. I., & Barto, A. G. (1991). Task decomposition through competition in a modular connectionist architecture. *Cognitive Science, 15,* 219–250.

Jacobs, R. A., Jordan, M. I., Nowland, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation, 3,* 79–87.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Keleman, D., & Bloom, P. (1994). Domain-specific knowledge in simple categorization tasks. *Psychonomic Bulletin & Review, 1,* 390–395.

Marechsal, D., & Schultz, T. R. (1996). Generative connectionist networks and constructivist cognitive development. *Cognitive Development, 11,* 571–603.

Markman, E. M. (1989). *Categorization and naming in children.* Cambridge, MA: MIT Press.

Medin, D. L., & Ross, B. H. (1997). *Cognitive psychology* (2nd ed.). Fort Worth: Harcourt Brace.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology, 19,* 242–279.

Mooney, R. J. (1993). Integration theory and data in category learning. In G. V. Nakamura, R. Taraban, & D. L. Medin (Eds.), *The psychology of learning and motivation: Categorization by humans and machines* (Vol. 29, pp. 189–218). San Diego: Academic Press.

Mozer, M. C., & Smolensky, P. (1989). Using relevance to reduce network size automatically. *Connection Science, 1,* 3–16.

Murphy, G. L. (1993). Theories and concept formation. In I. V. Mechelen, J. Hampton, R. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 173–200). London: Academic Press.

Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 904–919.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316.

Murphy, G. L., & Wisniewski, E. J. (1989). Feature correlations in conceptual representations. In G. Tiberghien (Ed.), *Advances in cognitive science* (Vol. 2, pp. 23–45). Chichester: Ellis Horwood.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101,* 53–79.

Peirce, C. S. (1931–1935). *Collected papers of Charles Sanders Peirce.* Cambridge: Harvard University.

Prechelt, L. (1997). Investigation of the CasCor family of learning algorithms. *Neural Networks, 10,* 885–896.

Raiffa, H., & Schlaifer, R. (1961). *Applied statistical decision theory.* Boston: Harvard University, Graduate School of Business Administration.

Reed, R. (1993). Pruning algorithms: A survey. *IEEE Transactions on Neural Networks, 4,* 740–746.

Schyns, P. G., Goldstone, R. L., & Thibaut, J. P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences, 21,* 1–40.

Spalding, T. L., & Murphy, G. L. (in press). What is learned in knowledge-related categories? Evidence from typicality and feature frequency judgments. *Memory & Cognition, 27.*

Vapnik, V., & Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications, 16,* 264–280.

Ward, T. B. (1994). Structured imagination: The role of category structure in exemplar generation. *Cognitive Psychology, 27,* 1–40.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12,* 129–140.

Wisniewski, E. J. (1995). Prior knowledge and functionally relevant features in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 449–468.

Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science, 18,* 221–282.