Background Knowledge and Models of Categorization

Evan Heit
University of Warwick

To appear in: U. Hahn & M. Ramscar (Eds.), Similarity and Categorization, Oxford University Press, 2000.

January 2000
10,500 words

Introduction

In most applications of formal models of categorization, category learning is portrayed as the building-up of a representation in memory for members of the category that have been observed. This assumption is perhaps the most basic that is made for models of categorization, that the representation of a category describes its observed members. Yet if category representations are to serve a purpose such as recognizing new members of a category, then simply relying on memory for known members would be a poor strategy in many situations. For example, if you are learning to distinguish the Smith family from the Jones family, and you have observed a tall, red-haired 45 year old woman who is the mother in the Smith family, and you then see another tall, red-haired 45 year old woman, you would probably classify her as belonging to the Jones family, despite her similarity to an observed member of the Smith family.

This example highlights the point that when few members of a category have been observed (or in the case of the Jones family, no members), it is crucial to rely on background knowledge rather than observed category members. This point is actually quite general, because there are many everyday situations where we are learning about new categories, such as visiting someplace new and learning about the social groups, buildings, landscapes, and so on, there. In each situation, starting with a fresh, or empty, category representation, and simply accumulating observations, would lead to great difficulties until a representative sample of category members can be observed. In the example of visiting a new place, background knowledge of people and buildings in other places would be crucial to classifying and reasoning in the new context, even if strictly speaking this information does not constitute observations of the target categories. That is, if you are trying to draw inferences about a novel category for which you have observed few category members or no category members, the best you can do is hope that this novel category will be like some previously known categories, and reason about the novel category based on knowledge of other categories.

A great deal of work (reviewed in the Ahn & Dennis and Hampton chapters in this volume) has contrasted the similarity-based approach to categorization and the knowledge-based approach. The similarity-based approach highlights the use of similarity to observed category members and the knowledge-based approach emphasizes other processes and use of

theoretical knowledge. However, in some ways this is a false dichotomy. Just because people are using information beyond the category members observed, the same sorts of similarity-based processing might be applied to observations from outside the immediate context. A model of similarity-based categorization could be extended to use category members from, say, outside of a psychology experiment. It is important to see how far similarity-based models can go, before postulating other sorts of processing (such as use of theories) which may not be as well-specified.

In fairness, most applications of models of categorization have not addressed the background knowledge issue (but for exceptions in psychological research, see Heit, 1997, for a review, and for a review of relevant artificial intelligence research, see Heit & Bott, 2000). Instead, most modeling work has addressed situations for which prior knowledge has little apparent effect, because the categories are well-learned or because there is not much prior knowledge that is relevant. But there is nothing in principle that keeps categorization models from addressing background knowledge effects. In this chapter, the main focus will be the integration model of categorization (Heit, 1994), which is one similarity-based model that has already been applied to a series of experiments on background knowledge effects (Heit, 1994, 1995, 1998).

This chapter addresses the generality of the integration model by applying it to a set of experiments designed by <u>other</u> researchers. How far can a similarity-based model of categorization go, applying ordinary assumptions about categorization and without making any special claims about other kinds of theoretical processing?

The Integration Model

The integration model is an exemplar model of categorization that has been extended to apply to situations in which previous knowledge has an influence. The most critical claim for the integration model is that when a person judges whether item X belongs to a newly-learned category, A, this item is both compared to actual observations for category A as well as being compared to a representation of prior knowledge for this category. For example, imagine that you are visiting Vienna for the first time, and you are trying learn about a local style of architecture called Jugendstil, which you have never observed before. However, you are aware that this category of buildings is similar to Art Nouveau, and you have seen many Art Nouveau buildings in Paris. According to the integration model, your representation of the Jugendstil category would summarize your observations of Jugendstil buildings (initially,

none) and also contain summary information about the related category, Art Nouveau. Classifying a building as Jugendstil in Vienna could be derived from its similarity to observed members of the Jugendstil category in Vienna or it could be based on its similarity to observed members of the Art Nouveau category in Paris. As you observe more Jugendstil buildings in Vienna, your representation of the Jugendstil category will increasingly rely on actual observations of category members and less on prior knowledge about Art Nouveau buildings. The integration model is closely related to accounts of judgment such as anchor-and-adjust procedures (Tversky & Kahneman, 1974) and Bayesian revision techniques (e.g., Edwards, Lindman, & Savage, 1963). The integration model itself is described formally in the Appendix.

Although the integration model may seem like a plausible account of some influences of background knowledge on category learning, it is by no means the only possible account. Heit (1994, 1997, 1998) has reviewed several other possibilities. For example, prior knowledge might have a selective weighting influence (Keil, 1989; Murphy & Medin, 1985). By this account, learners might especially attend to features, configurations of features, or observations that are congruent with prior knowledge, and ignore information that does not fit with previous knowledge. (Or alternately, selection could be in favor of observations that contradict prior knowledge, cf., Heit, 1998.) Although previous work (Heit, 1993, 1994) has compared the integration and weighting accounts, this chapter is not intended for the most part to directly compare the integration model to other accounts. The experiments analyzed here were not designed to distinguish the integration model from other possible accounts. Rather, the present goal is to illustrate how the integration model can be applied to a range of studies.

Overview of Experiments to be Analyzed

This chapter will present simulations of seven experiments from four published papers. These experiments are diverse enough to cover a range of possible techniques. Wattenmaker, Dewey, Murphy, and Medin (1986), Experiment 3, taught adult subjects about categories of people in different occupations, using stimuli consisting of feature lists. The procedure was to train subjects on classification until a learning criterion was reached. Pazzani (1991), Experiment 1, was similar in these methodological elements except that the stimuli were color photographs of people inflating balloons. Barrett, Abdi, Murphy, and Gallagher (1993), Experiments 1 and 2, tested children below age 10, using picture stimuli

describing kinds of birds and other animals as well as tools. Their procedure involved separate training and transfer phases. Finally, Murphy and Allopenna (1994), Experiments 1, 2, and 3, also used separate training and transfer phases, on adult subjects, with feature list stimuli describing kinds of animals, buildings, and vehicles.

Furthermore, these experiments had a number of different aims. The Wattenmaker et al. and Pazzani studies were looking at the influence of prior knowledge on learning particular category structures. To what extent does accessing background knowledge make learning some category structure easier or more difficult? Barrett et al. were concerned with how prior knowledge helps children to attend to feature correlations that will be useful for category learning, on the assumption that it would be difficult for children to consider all possible combinations of features when learning (cf., Murphy & Medin, 1985). Murphy and Allopenna addressed several issues, including what kinds of knowledge will facilitate category learning and whether prior knowledge will ever have harmful effects on learning, i.e., diminishing sensitivity to what is actually observed in the category.

These experiments were analyzed using a formal model of categorization, implementing the integration account. The modeling is described in detail in the Appendix, and only a few points are emphasized here. In implementing the integration model, it was assumed that categorization judgments are influenced by similarity to prior examples and by similarity to observed category members. The critical free parameter in the integration model is $\underline{k}$, which refers to the relative influence of prior examples on categorization. The $\underline{k}$ parameter may vary from 0 to 1, and the quantity $(1 - \underline{k})$ indicates the relative influence of actual observed category members. Also, some assumptions had to be made about the nature of the prior examples themselves—what prior knowledge are people drawing on? The issue of how people select prior knowledge that will be useful for category learning was not addressed by these simulations, however this issue will be returned to in the General Discussion. For now, plausible assumptions will have to be made about prior knowledge retrieved, to show how the integration model might be applied.

The stimuli in the experiments were described in terms of multiple dimensions, such as color, size, and shape. In the models, the attention to each stimulus dimension is indicated by the value of a $\underline{w}$ parameter. These parameters may vary from 0 to 1, with lower $\underline{w}$ values indicating greater attention to a dimension. For example, if subjects are influenced more by color and size compared to shape, that would be represented by lower $\underline{w}$ values for color and size than for shape.

The simulations for each experiments or group of experiments will be presented in two sections. First there will be a summary of the main findings. Then there is the more detailed presentation of the simulations themselves, for the interested reader.

Wattenmaker et al. (1986), Experiment 3

Key Findings

It is plausible that certain kinds of category structures might be easier or more difficult to learn depending on the prior knowledge that is accessed. Medin and Schwanenflugel (1981) distinguished between two kinds of classification structures, linearly separable and nonlinearly separable. If a pair of categories, A and B, are linearly separable, then by definition it is possible to classify a new stimulus, x, using a simple linear rule. One such linear rule would be to count whether x has more characteristic features of category A or of category B. In contrast, if A and B overlap to the extent that they are nonlinearly separable, then no linear rule will allow perfect discrimination between members of the two categories. (However, a nonlinear rule that considers configurations of features in addition to single features could be successful.) In some sense, a linearly separable structure seems easier to learn. Wattenmaker et al. addressed the issue of whether prior knowledge can help the learning of nonlinearly separable (NLS) structures compared to linearly separable (LS) structures. If prior knowledge suggests an NLS structure, then this kind of categorization task should be easier.

The subjects in this experiment learned to distinguish between a category of house painters and a category of construction workers. Wattenmaker et al. assumed that usually subjects would expect house painters to work indoors, work in a small crew, and work year-round, and expect construction workers to work outdoors, work in a large crew, and to not work in the winter. Subjects in the Uninformed conditions were not provided with further information, but subjects in the Informed conditions were given a hint to consider that house painters can work indoors or outdoors. Informed subjects were expected to look for two feature configurations: house painters who work indoors all year-round and house painters who work outdoors and do not work during the winter. The Informed subjects were expected to be facilitated at learning an NLS structure, and impaired at learning an LS structure, due to their expectations about a configuration of features. Table 1 shows that these results were obtained, in terms of overall average errors during learning.

The integration model was applied by assuming that, in addition to observed category members, people responded to similarity to a number of prior examples, derived from background knowledge. In applying the model, the construction worker category had one prior example, corresponding to the usual stereotype described above. In the Uninformed conditions, the house painter category had one prior example, corresponding to the stereotype described above. In the Informed conditions, the simulation was provided with two prior examples for this category, corresponding to an indoor house painter and an outdoor house painter. With these assumptions, the model predicts the key result, that the NLS structure is easier than the LS structured, for Informed subjects. At a finer level of description, the correction between the model's prediction and the data was fairly good, .84.

Simulation Details

The model was fitted to the mean number of errors made on the eight stimuli in each condition. Because Wattenmaker et al. did not collect transfer data, their results on errors during learning were taken as a proxy for a hypothetical transfer phase. It was assumed that if some item had a low number of errors during training, then subjects would likely be especially accurate on this item during a transfer test as well. It would be possible to model training data directly (e.g., Estes, Campbell, Hatsopoulis, & Hurwitz, 1989), but additional information would be needed, such as the error rate on each item in each block of training. One complication is that the influence of prior knowledge could vary over the course of learning (Heit, 1994). Also, the accuracy in encoding stimuli, as expressed by the w parameters, might vary across blocks (Taraban & Palacios, 1993).

The integration model was fitted by maximizing the correlation between the number of learning errors on the items and the error rates predicted by the model. Note that in Table 1, the integration model predicts the correct pattern at the level of averages, that there are more errors in LS-Informed than in LS-Uninformed, and that there are fewer errors in NLS-Informed than in NLS-Informed. The correlation between the integration model's predictions and the data was .84, with the relative influence of prior knowledge, k, estimated to be .40 in the Uninformed conditions and .18 in the Informed conditions. (The lower value of k for the Informed conditions may be attributed to the higher number of prior examples used.) The attention weights were estimated to be .32, .27, and .41 for the LS structure and .09, .00, and .24 for the NLS structure. The performance of the integration model can be considered as quite reasonable, considering that it captures the overall results of this

experiment, and it was fitted to error-rate training data that are perhaps less than ideal for modeling.

Pazzani (1991), Experiment 1

Key Findings

In an elegant study, Pazzani (1991) investigated the issue of knowledge effects by teaching subjects about categories of balloons, an already-familiar kind of object. Similar to Wattenmaker et al. (1986), Pazzani addressed the interaction between prior knowledge and kind of category structure to be learned, comparing two forms of linearly separable structures, conjunctive and disjunctive. Past research (Bruner, Goodnow, & Austin, 1956) suggested that conjunctive structures, defined by a set of features in an "and" relation, are easier to learn than disjunctive structures, defined by a set of features in an "or" relation.

In Pazzani's Experiment 1, subjects were instructed either to learn about category of balloons that inflate or to learn about category that was simply labeled "Alpha." It was assumed that subjects in the Inflate conditions would be influenced by their prior knowledge of what it takes to inflate a balloon, whereas prior knowledge would not affect the Alpha condition. A pretest showed that people expected that stretching a balloon would facilitate inflation and that adults would be more successful than children at inflation. The stimuli in this experiment were pictures of scenes that varied on four dimensions: adult or child, stretched balloon or balloon dipped in water, yellow or purple balloon, and small or large balloon. In the Disjunctive conditions, the Inflate (or Alpha) category was defined by a disjunctive rule: These balloons must be stretched or inflated by an adult. Note that this disjunctive rule is consistent with subjects' expectations about inflating balloons. In the Conjunctive conditions, the target category was defined by a conjunctive rule: these balloons must be small and yellow. Note that these characteristics are not consistent with people's expectations about inflation.

As anticipated, Pazzani found that learning was faster in the Disjunctive-Inflate condition (9.4 trials on average to a learning criterion) than in the Disjunctive-Alpha condition (30.8 trials). Also, learning was slower in the Conjunctive-Inflate condition (29.1 trials) than in the Conjunctive-Alpha condition (18.0 trials).

Because balloons are familiar objects, it is highly plausible that subjects in this experiment also remembered balloons they had seen in other contexts. Similarity to these extra-experimental observations could also influence performance in this experiment, in

addition to similarity to the category members observed within the experiment. To evaluate this proposal, the integration model was simulated with just a single prior example, of an adult stretching a balloon, placed in the Inflate category. With this simple assumption made, that subjects remembered seeing an adult stretching a balloon then inflating it, the integration model predicts the correct pattern of results, in terms of the order of difficulty for the different conditions.

Simulation Details

Because of the small number of independent data points in this experiment, the model was not fitted directly to the data, but rather, plausible parameter values were chosen to demonstrate the qualitative predictions of the models. The relative influence of prior knowledge, k, in the Inflate conditions was set to .98. This large influence of prior knowledge leads the integration model to predict that subjects would reach the criterion quickly in the Disjunctive-Inflate condition and learn very slowly in the Conjunctive-Inflate condition. The attentional weight parameters, w, were set to .15 for all features of the positive observations (inflatable balloons or Alphas), and the w values were all set to .80 for the negative observations (non-inflatable balloons or non-Alphas). In other words, it was assumed that subjects paid more attention to positive instances than to negative instances. Unlike the other experiments described in this paper, the Pazzani experiment used an asymmetrical task in which subjects were instructed to learn a single category rather than to learn to distinguish between two categories. The assumption of different values for positive and negative instances allows an exemplar model of categorization to predict the result that Conjunctive-Alpha will be easier to learn than Disjunctive-Alpha. However, this assumption of asymmetry is not needed to predict the general trends of the prior knowledge effects in this experiment. In particular, even if the same attentional weights are used for positive and negative instances, the integration model still correctly predicts that prior knowledge will help learning in the Disjunctive conditions and impair learning in the Conjunctive conditions.

The integration model predicts that subjects would make fewer errors on transfer trials in the Disjunctive-Inflate condition (an average of .24 per item) than in the Disjunctive-Alpha condition (.47 errors). Also, the integration model predicts more errors in the Conjunctive-Inflate condition (.46) than in the Conjunctive-Alpha condition (.31). In other words, the integration model correctly predicts that prior knowledge will be helpful in the Disjunctive conditions and harmful in the Conjunctive conditions. This simulation that the results of Pazzani (1991) are consistent with the facilitation effects predicted by the

integration model. The fit of the integration model depends on the assumption that people have memories of adults stretching balloons and then successfully inflating them. This assumption seems intuitive and it is also consistent with the results of Pazzani's pre-test. Of course, it is quite plausible that people have even more elaborate knowledge about balloons, but the point is that the actual results of the experiment follow from these basic assumptions about what people know.

Pazzani's paper is quite valuable because it includes not only experimental data but computer simulations as well. Pazzani applied a rule-based categorization model, known as POST HOC, to his data. A critical similarity between the integration model and POST HOC is that in both models, an initial knowledge representation is revised as new observations are encountered. In the POST HOC model, this knowledge representation is encoded as rules, such as "age = adult -> inflate." On the basis of the experimental results, it would seem difficult to distinguish between memories of adults inflating balloons (as assumed by the integration model) and knowledge of rules such as adults are likely to be able to inflate balloons (as assumed by POST HOC). Indeed, it seems extremely likely in this case that people have both the example-based knowledge and the rule-based knowledge. The memories of the particular examples could have led to the rule learning, and knowledge of the rules could support the memories for examples. In sum, the results of Pazzani's experiment are compatible with either model, and furthermore it may be the case that two models are complementary to each other.

Barrett et al. (1993)

Key Findings

The next set of experiments, by Barrett et al. (1991), used children as subjects, at an age where they are learning about many everyday categories at school and in their daily observations. The two experiments were particularly intended to focus on the issue of selective attention, that is, does children's prior knowledge help them to focus on some feature correlations and ignore others—the role of prior knowledge highlighted by Murphy and Medin (1985) (see also Murphy and Wisniewski, 1989). Although it is extremely likely that knowledge can help people to select features, the question is what actually serves as distinctive evidence for feature selection. Notably, the integration model gives an alternative account, in terms of prior examples being put together with observations of category members.

For example, in Experiment 1, children learned about descriptions of birds. In a classification test, performance was better for items that maintained an expected correlation (e.g., small brain and poor memory) compared to items that broke the expected correlation (e.g., small brain and good memory). The integration account provides a ready explanation for this result. In addition to the observed category members, children should be able to recall information from outside of the experiment, such as animals with small brains and poor memory, and animals with big brains and good memory. The integration model predicts that, due to the influence of these prior examples, it should be easier to classify a new item that maintains the correlation compared to an item that violates the correlation.

In Experiment 2, Barrett et al. taught children about two categories of animals (or two categories of tools). The categories had fictional names, but one category had a lot in common with mountain animals and the other category had a lot in common with desert animals. The test stimuli were equally similar to what had been studied for both categories, but they were more similar to the stereotype (according to prior knowledge) for one category. For example, a test stimulus might have two features in common with what had been studied for category 1 and two features in common with what had been studied for category 2, but the test stimulus was also similar to the stereotype for mountain animals (it has wool). Children were likely to place this test item in category 1. This result fits with the assumption of the integration model that children retrieve prior examples, of mountain animals and desert animals, which also influence judgments about category 1 and category 2, respectively.

Simulation Details

Experiment 1. In the training phase of this experiment, children learned about two categories of birds, with unfamiliar labels, jasslers and loppets, presented as drawings. Each category member had three of the four features contained in a prototype. The prototype of one bird category was: big brain, good memory, two-part heart, and rounded beak, and the prototype of the other bird category was: small brain, poor memory, three-part heart, and pointed beak. Note that the first two dimensions, brain size and memory, are related according to general knowledge, such that some animals would be expected to have larger brains and better memories, and other animals would be expected to the have the contrasting features.

In the test phase, the Correlation items maintained the correlation between knowledge relevant-features (e.g., a test item had a big brain and a good memory), and the Broken-correlation items violated the correlation (e.g., had a big brain and a poor memory). These

critical test items each contained two or three features of one category and one feature from the contrasting category. Barrett et al. found that children placed the Correlation items in the target category on 75% of the trials, and placed the Broken-correlation items in the target category on only 48% of the trials. This pattern of results can be explained readily in terms of the integration account, without resorting to selective attention to particular features. By the account of the integration model, children retrieved a prior example of large-brained animal with a good memory for one category and a small-brained animal with a poor memory for the other category. With the k parameter set at .8, and the w parameters all set at .5, the integration model predicts that subjects would make the target classification on 80% of Correlation items but only on 53% of Broken-correlation items. In sum, the integration model can capture the key result of this experiment, but with a different interpretation that Barrett et al., without assuming selective attention to particular feature correlations.

Experiment 2. Children in the second experiment were instructed to either learn about two categories of animals or about two categories of tools. The training stimuli were the same for both conditions; the descriptions were crafted so that they could be either of animals or tools. For the purpose of constructing training stimuli, one category had this prototype: found in the mountains, has wool, crushes rocks, shaped like an arrow, and catches snakes. The prototype of the other category was: found in desert, stores water, cuts cactus, shaped like a plate, and catches spiders. According to a pretest, subjects had strong prior expectations about the first and second dimensions in the Animal condition (e.g., they expected some animals to live in the mountains and have wool), and subjects had strong prior expectations about the first and third dimensions in the Tool condition (e.g., some tools would be found in the mountains and be used to crack rocks).

The pattern of results for Experiment 2 was similar to Experiment 1; subjects were especially influenced by the knowledge-relevant features. A number of test stimuli had two knowledge-relevant features from one category and two knowledge-irrelevant features from another category. If subjects were not influenced by prior knowledge, their classifications would be at chance for these stimuli. In the results reported by Barrett et al., these stimuli were classified approximately 80% of the time in the category for which the stimulus had knowledge-relevant features. The integration model was fitted to results for the 16 transfer items reported by Barrett et al., using a log-likelihood criterion, and the model yielded a high correlation, r = .96, and a low average error, RMSE=.1011 (log-likelihood –24.41). The estimated influence of prior knowledge, k, was .38, and the estimated w parameters for the

five dimensions were .07, .60, .46, .01, and .99. Note that the dimension weights were kept exactly the same for the Animal and Tool conditions. Therefore these weights cannot be sensitive to different prior knowledge in the two conditions. The model captures the results of individual test items fairly well, as shown in Table 2. Unfortunately for the purpose of the modeling, this paper did not report the complete set of results for transfer stimuli. Performance of the model could be well better on a more complete data transfer set. Again, however, as in Experiment 1, the integration model gives a good account of the data without resorting to prior knowledge effects in terms of selective attention.

Murphy and Allopenna (1994)

Key Findings

These experiments were intended to investigate various knowledge-related influences, including meaningfulness and familiarity, on category learning. The categories were labeled simply as Category 1 and Category 2. The prototype for each category had five characteristic features, and each category member was described with three of these features. In addition, each category member had a pair of random features selected from a set of features that appeared about equally often in two contrasting categories. The main logic of these experiments was to keep the structure of the categories the same but vary their content, i.e., the nature of their features. For example, the features might be typographical symbols or they might be descriptive words. With the descriptive word stimuli, the categories varied in how well they matched up with prior knowledge. For example, in the Meaningful conditions, the stimuli were not actually too meaningful: A sample description would be made in Africa, has barbed tail, fish kept as pets. In the Integrated conditions, the stimuli were more like those of Barrett et al. (1993); they matched up with familiar sub-categories such as jungle vehicles and Arctic vehicles. Murphy and Allopenna found better performance in the Meaningful conditions compared to using typographical symbols, and better still performance in the Integrated conditions. However, sensitivity to frequency manipulations in the Integrated conditions, for what had actually been presented for the categories, was poor.

The general way for the integration model to address these results would be to assume that people retrieve prior examples of known types of vehicles that also affect judgments, particularly in the Integrated conditions. For example, people might rely on memories of jeeps and Land Rovers for the category of jungle vehicles. This supplementary use of prior example would predict good overall performance in the Integrated conditions, e.g.,

distinguish jungle vehicles from Arctic vehicles. But they heavy use of prior examples would also predict low sensitivity to frequency manipulations for what actually had been presented, as if people were relying much more on stereotypes than on observations. Although the integration model correctly predicts that performance will be best overall in the Integration condition, it does not predict that performance would be worst overall for typographical symbols. Other assumptions would be needed to explain why memory is poor for typographical symbols.

Simulation Details

Experiment 1. Subjects in the three conditions of this experiment saw stimuli that were structured in the same way, but the features themselves differed considerably. In the Arbitrary condition, the features were typographical symbols, so that the prototype for one category might be +, {, >, $, and [. In the Meaningful condition, the individual features were meaningful but they were completely unrelated to each other. For example, the prototype for one category was lives alone, made in Africa, fish kept as pets, has barbed tail, and has thick heavy walls. These features were taken from different kind of real-world categories, including animals, buildings, and vehicles. Finally, in the Integrated condition, the features were both meaningful and coherently related. For example, the prototype for one category was made in Africa, lightly insulated, green, drives in jungles, and has wheels, that is, all from the same domain, vehicles. It was expected that in this condition, the category members would elicit prior knowledge of familiar categories such as vehicles. The random features in the Integrated condition were not related to the features of the prototypes. The critical result was the number of training blocks it took subjects to reach a learning criterion. As shown in Table 3, subjects learned most slowly in the Arbitrary condition and most quickly in the Integrated condition.

Insert Table 3 about here

The integration model can explain the advantage of the Integrated condition over the Meaningful condition, by assuming that subjects in the Integrated condition retrieved appropriate prior examples with the knowledge-relevant features. In other words, it is assumed that people can assemble a representation such as “jungle vehicles” based on prior knowledge. Table 3 shows the predicted error rates from the integration model, with the assumption that k is .5 for the Integrated condition and zero for the other conditions, and w is

.5 for each stimulus dimension. The integration model, however, does not predict a difference between the Meaningful and Arbitrary conditions, because the incoherent category members in the Meaningful condition would not elicit prior examples. There seems to be an additional benefit of using meaningful stimuli rather than the typographical symbols, not captured by the integration model.

Experiment 2. This next experiment included the Integrated and Meaningful conditions again. In addition, a new, infrequent stimulus feature was introduced. One category member contained an infrequent feature that was not present in any other category member. In contrast, the frequent features appeared in six or seven training items. The question of interest was whether reliance on prior knowledge in category learning could ever be detrimental to learning the finer details of categories, such as information about frequency of rate features. For the Integrated condition, the infrequent feature was intended to be related, according to prior knowledge, to the other features. For example, the infrequent feature for the jungle vehicle category was "convertible." Again, in the Meaningful condition the features were drawn from different, incoherent domains.

Table 3 shows the most critical results, for transfer test items that either had a Frequent feature or an Infrequent feature. In the Meaningful condition, the error rate is lower for the Frequent items, which is not surprising considering their greater presentation frequency during training. The more striking result is in the Integrated condition, where subjects responded the same to Frequent and Infrequent items. It appears that judgments on Infrequent items were greatly affected by prior knowledge, considering that there was only one observation of each infrequent feature during the training phase. (See Heit and Bott, 2000, for a related result in which sensitivity to frequency is quite low in the face of knowledge-driven category learning.)

By assuming that the prior examples include information about the Infrequent features, the integration model can predict this interaction between Meaningful versus Integrated condition and Frequent versus Infrequent test item. For example, if prior knowledge can be used to infer that jungle vehicles are likely to have an open top, then subjects should classify this feature correctly even if it is only presented rarely. The predictions are shown in Table 3 for $w = .1$, and $k = .7$ in the Integrated condition. According to the integration model, accuracy is high for Infrequent items because they match the prior examples well.

Experiment 3. The third experiment was similar to Experiment 2, except that a Domain condition was added. In this condition, the stimulus features were coherent in that they all came from the same domain. For example, one category was constructed from variants of this prototype: green, manual, radial tires, air bags, vinyl seat covers, and convertible. This is a sensible object, unlike the prototypes in the Meaningful condition which included features from unrelated domains such as parts of animals, vehicles, and buildings. However, the features in the Domain condition were not interpredictive as in the Integrated condition, which had features such as made in Africa, lightly insulated, and driven in jungles. So prior knowledge of existing categories would not help to draw inferences about the features of this new category. The results of the training phase are shown in Table 3. The learning rate in the Domain condition was intermediate between the Meaningful and Integrated conditions, but the difference between the Domain and Meaningful conditions did not reach statistical significance.

For the integration model, it was assumed that prior examples would be used in the Integrated condition ($k = .5$), but that k would be zero in the Domain and Meaningful conditions. The w parameter was set at .50 for each dimension. The predictions are shown in Table 3. The predictions of the integration models is consistent with the results, considering that it is difficult to say whether there is truly a difference between the Meaningful and Domain conditions.

In a posttest conducted by Murphy and Allopenna (1994), subjects were asked whether the prototypes for the Domain and Integrated conditions corresponded to a real, familiar object. The Domain prototypes were rated as familiar objects on 64% of test trials, and the Integrated prototypes were rated as familiar on only 24% of test trials. In Experiment 3, learning was faster in the Integrated condition than in the Domain condition, yet the posttest showed that the Domain category members were more familiar. This result suggests that the learning of a category will be facilitated when it is compatible with prior knowledge of how dimensions are interpredictive, as in the Integrated condition. In the case of jungle vehicles, it is helpful for new categories to fit with prior categories that may be quite rare. In contrast, being compatible in a general sort of way with a lot of familiar knowledge may not help learning. The integration model must assume that interpredictivity between features, rather than just familiarity, is critical to retrieving prior examples. Although subjects may have never seen a vehicle with all of the characteristics of the Integrated prototype, they probably knew of real objects that preserve some of the predictive relations between the features. For example, they might have prior examples of lightly-insulated jungle buildings,

green clothing in jungles, jungles in Africa, and convertibles in warm climates. Murphy and Allopenna's own explanation for their results is that knowledge facilitates the identification or construction of a schema during learning, and that subsequent judgments are influenced by similarity to the schema (p. 916). Likewise, the integration model must assume that the prior examples are retrieved and assembled during learning, and these prior examples affect subsequent judgments.

The final result from Experiment 3 is that on a transfer test, subjects were least sensitive to feature frequency in the Integrated condition. That is, subjects responded nearly as strongly to Infrequent items as to Frequent items. As explained for Experiment 2, the integration model readily predicts this result.

General Discussion

Integration and Knowledge Selection

The analyses presented in this chapter have been informative in showing to what extent background knowledge effects in categorization can be attributed to two kinds of similarity information: similarity to observed category members and similarity to prior examples derived from other contexts. The prior examples assumption of the integration model is perhaps somewhat controversial. One difficulty in assessing the integration model is that, unlike the training stimuli and subjects' responses in an experiment, prior examples are by their nature unobservable, because they precede the experimental context. A further difficulty in studying prior examples is that it is natural to assume that people are flexible in their use of prior knowledge. For example, in Barrett et al. (1993), Experiment 1, subjects expected that birds with large brains would have superior memory abilities. Yet it is possible that subjects only previously knew of small-brained birds. The expectations for large-brained birds could have been derived from prior examples of other kinds of large-brained animals, e.g., humans (cf., Carey, 1985). (See the Rodriguez chapter in the volume for further discussion on how examples may be adapted, in the context of case-based reasoning systems.) In future evaluations of the integration model, it may be useful to ask subjects for information about their prior examples, as in Murphy and Allopenna (1994). However, such enquiries should be interpreted with caution, because a question from an experimenter is only an indirect assessment of subjects' knowledge; it does not constitute a direct observation of the prior examples that the subjects might use during categorization.

One important aspect of knowledge use in category learning, not captured by the integration model, is that people need to figure out which knowledge will be useful to support learning. Particularly for the Barrett et al. (1993) and Murphy and Allopenna (1994) studies, the category labels did not give any guidance about what the content of the categories. Subjects had to actually observe some category members before deciding which knowledge would be helpful, or in the case of some of Murphy and Allopenna's stimuli, the subjects probably failed to retrieve any helpful prior knowledge. The mechanisms involved in knowledge selection were not part of the simulations here, however, knowledge selection could possibly be built into an exemplar model. For example, in Barrett et al.'s Experiment 1, observing a small-brained, unintelligent bird during training could lead to a reminding of a similar bird that was observed prior to the experiment. Heit (1992) developed a model of categorization that uses such chains of reminding, and the assumptions of the Heit (1992) model could be incorporated within the integration model to describe the reminding process more fully. (See Ross, Tenpenny, and Perkins, 1990, for further evidence of the influence of remindings during category learning.)

Recently, Heit and Bott (2000) have addressed the knowledge selection problem in experiments that were variants of Murphy and Allopenna (1994). In these new experiments, subjects learned about novel categories of buildings and vehicles, but the categories were initially given uninformative labels (the Doe and Lee categories). Because people know about a very large number of building and vehicle categories, it was difficult to form advance expectations about the novel categories—too many possible sources of prior knowledge might apply. Unlike Murphy and Allopenna, Heit and Bott collected data over the time course of learning, that is, after various numbers of category members had been observed. This technique gave a more complete picture of how knowledge was used at different times, and also produced a data set that was better suited to formal modeling. The key result was that initially, prior knowledge did not affect subject's judgments, but its use built up over time. More concretely, some subjects learned about the Doe category of buildings, which resembled churches, and the Lee category of buildings, which resembled office blocks. These buildings were described in terms of critical features related to these themes, such as steep roofs and wooden benches for Does and flat roofs and metal furniture for Lees. Also there were filler features that could be true for any kind of building, such as near a river or designed by a local architect. In the early stages of learning there was no advantage for critical features over filler features, because subjects simply had not realized that churches were relevant to Does and office buildings were relevant Lees. However, with more

observations, there was an increasing advantage for critical features (see Figure 1 for representative results).

Heit and Bott (2000) developed a computational model, the Baywatch model, that addresses some aspects of knowledge selection. Although this new account is a neural network model rather than an exemplar model, it has much in common with the integration model. In particular, the Baywatch model derives its classification judgments by summing up information from actual observed category members and from information obtained from past categories. However, the Baywatch model also has a mechanism for selecting from among different sources of prior knowledge. In effect, already known categories, such as known categories of buildings, compete with each other to make predictions about new categories. In the present example, prior knowledge about churches is eventually successful in winning the competition to make predictions about Does, and likewise the network learns to use knowledge about offices to make predictions about Lees (see Figure 2 for representative predictions of the model).

Other Results and Processes

Other studies of categorization have uncovered additional influences of prior knowledge, implicating processes in addition to integration. These studies are important but no attempt was made here to model their results, because they fall out of the boundaries of the simulations in this paper. Still, complete account of effects of knowledge on categorization ultimately would have to address these phenomena.

First, prior knowledge may influence how observed category members are encoded. (Heit, 1993, 1994, referred to such influences as distortion processes.) For example, in Experiment 4 of Wattenmaker, et al. (1986), subjects were encouraged to interpret various activities as sports versus nonsports, or as indoor sports versus outdoor sports. So, a person who enjoys baseball might be encoded either as enjoying a sport or enjoying an outdoor sport. This experimental manipulation had an impact on subjects' ease of learning categories, but it is not well-explained in terms of integration, selection, or facilitation processes. Wisniewski and Medin (1994) provide a more extensive discussion of feature interpretation processes. (See also Schyns, Goldstone, & Thibaut, 1998.) Wisniewski and Medin used experimental stimuli in the form of children's drawings. They found, for example, that part of a picture might be interpreted as either a pocket or a purse depending on subjects' expectations, such as whether the picture was drawn by a country child or a city child. Likewise, research on scientific and medical expertise has shown that knowledge has

pervasive influences on how new category members (e.g., physics problems and x-rays) are interpreted (Chi, Feltovich, & Glaser, 1981; Lesgold, Rubinson, Feltovich, Glaser, Klopfer, & Wang, 1988).

A second additional influence of knowledge on categorization is that knowledge may sometimes affect how similarity information is used. Lamberts (1994) taught subjects about categories, or families, of schematic faces. The transfer stimuli in this study were described as either brothers or cousins of the original training items. Lamberts inferred that subjects used different similarity functions to categorize brothers and cousins. Subjects used their general knowledge that siblings in a family may be close matches for each other, so subjects used a steeper gradient of generalization when categorizing brothers compared to categorizing cousins. Note that this result cannot be explained in terms of selective attention to particular features.

Conclusion

The integration model has been implemented as an exemplar model of categorization, but its use of exemplar representation is not a critical aspect of the model. Of course, the integration model is compatible with previous exemplar models (e.g., Estes, 1994; Heit, 1992; Lamberts, 1994; Medin & Schaffer, 1978; Palmeri, this volume) that have been widely applied. But the key psychological principle embodied by the integration model, that categorization depends on similarity to what has been observed as well as similarity to a representation of prior knowledge, could easily be implemented outside the framework of exemplar models. For example, Heit (1993) described a prototype-model version of integration, and Heit (1995) described a simple connectionist network that acted quite similar to the integration model. Within the framework of exemplar models, it is perfectly reasonable to describe prior knowledge in terms of prior examples. That is, it would be reasonable to treat background knowledge about buildings in terms of memories of prior exemplars of this category. But within other modeling frameworks, it would be just as reasonable to think of prior knowledge as prior prototypes, prior connection strengths, prior schemas, or prior production rules.

The present paper is not intended to distinguish among different possible forms of category representation. Although it is an interesting question to consider the format of background knowledge (e.g., is it instance-based, or does it has a distributed representation, or is it theory-like?), existing data simply do not answer this question. For example, different versions of the integration model, with exemplar, prototype, or network representations,

would have the same successes and failures in fitting the results considered here. To be specific, whereas the simulations in this chapter used exemplar representation for describing prior knowledge in terms of prior examples, similar predictions could be derived from other ways of representing prior knowledge, such as prototypes. Although representation is a crucial issue for categorization researchers, it is important to not presume that every study says something, or even ought to say something, about representation. Barsalou (1990) highlighted the dangers of claiming that particular experimental results indicate a particular kind of representation, because the predictions for a kind of representation depend heavily on the processing assumptions made. Rather than addressing representational questions, the present chapter is concerned with processing questions. In particular, how well do the processing assumptions embodied by the integration model explain this varied set of results?

It is useful to step back from the experiments described in this chapter to consider more broadly the phenomenon of background knowledge influencing category learning. This chapter has focused on different processes by which prior knowledge and observations can be put together, highlighting processes such as integration, feature selection, and selection of previous categories. Another facet of this issue, explicitly not addressed in this chapter, concerns the format of category representation, e.g., exemplars, prototypes, schemas. A third facet, which perhaps deserves increased attention in the future, is the content of prior knowledge. That is, what different kind of information are contained in the prior knowledge that guides category learning?

It is possible to propose a preliminary taxonomy of prior knowledge, ranging from specific to general information. Most of the examples in this chapter have addressed very specific prior beliefs that are incorporated into new categories, such as that Jugendstil buildings will be colorful or the target category of housepainters will work indoors. Next on the scale of generality would be beliefs that are more relational or need to be adapted before they can be applied to a new category. Wisniewski (1998) provides some good examples of how properties might be adapted in the service of learning about new categories (his focus was a slightly different topic, the interpretation of combined concepts). For example, imagine that someone is learning about the “giraffe duck” category. The expectations would probably would not involve a direct transfer of specific beliefs about giraffes, such as having an eight-foot-high neck. Rather, it would be expected that the duck’s neck would be long relative to other ducks. So the prior belief about long necks would be treated as a belief that had to be translated in order to serve as a belief for the novel category.

Third, prior beliefs could suggest an abstract category structure such as linearly separable or non-linearly separable categories (see also Wattenmaker, 1995). It could be the case that people expect linearly separable categories, for example, in some domains, such as social categories in general. Heit and Bott (2000) reviewed several ways that researchers have built constraints into neural networks that solve classification problems, and many of these schemes are at the abstract structural level. For example, in constructivist networks (e.g., Mareschal and Schultz, 1996), the system starts off with a small number of hidden units, constraining the network to learn simpler category structures. If applying a simple category structure is not successful, then more hidden units can be added, permitting more complex category structures. Hence, these networks have an implicit ordering of preferred category structures.

Finally, the beliefs could be so general that, in Goodman's (1955) terms, they would be over-hypotheses, that is, beliefs about beliefs. These would be beliefs that are so general that they tell people how to learn about categories. For example, according to the shape bias (e.g., Ward, 1993), the shape of an object is particularly informative about its category membership. The shape bias could guide category learning in rather unfamiliar situations, even when there are not any more specific prior beliefs to be applied. Another example would be general beliefs about how category members are distributed. For example, do several observations of some item count as one category member observed several times, or as several category members each observed once (Barsalou, Huttenlocher, & Lamberts, 1998)? Do people expect to see typical category members before atypical category members (Avrahami et al., 1997)?

Returning to the point at the start of this chapter, it should be clear when evaluating the statement that categorization depends on similarity, it is crucial to consider similarity to what. In addition to employing similarity to observed category members, models of categorization can use information about similarity to prior knowledge. The applications of the integration model in this chapter, to a variety of experiments, show that this is possible. In this way, many "knowledge effects" on categorization can be subsumed within similarity-based categorization, rather than being taken as evidence against similarity-based categorization.

Acknowledgements

This research was supported by a grant from BBSRC. Address correspondence to Evan Heit, Department of Psychology, University of Warwick, Coventry CV4 7AL, UK; E.Heit@warwick.ac.uk.

References

Avrahami, J., Kareev, Y., Bogot, Y., Caspi, R., Dunaevsky, S., & Lerner, S. (1997). Teaching by examples: Implications for the process of category acquisition. Quarterly Journal of Experimental Psychology: A, 50, 585-606.

Barrett, S. E., Abdi, H., Murphy, G. L., & McCarthy Gallagher, J. (1993). Theory-based correlations and their role in children's concepts. Child Development, 64, 1595-1616.

Barsalou, L. W. (1990). On the indistinguishability of exemplar memory and abstraction in memory representation. In T. K. Srull, & R. S. Wyer (Eds.), Advances in Social Cognition (pp. 61-88). Hillsdale, NJ: Erlbaum.

Barsalou, L. W., Huttenlocher, J., & Lamberts, K. (1998). Processing individuals in categorization. Cognitive Psychology, 36, 203-272.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). A Study of Thinking. London: Chapman & Hall.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5, 121-152.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. Psychological Review, 70, 193-242.

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

Estes, W. K., Campbell, J. A., Hatsopoulis, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 556-571.

Goodman, N. (1955). Fact, fiction, and forecast. Cambridge, MA: Harvard University Press.

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

Heit, E. (1993). Modeling the effects of expectations on recognition memory. Psychological Science, 4, 244-252.

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. Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society. 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. (1998). Influences of prior knowledge on selective weighting of category members. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20, 712-731.

Heit, E., & Bott, L. (2000). Knowledge selection in category learning. In D. L. Medin (Ed.), Psychology of Learning and Motivation, (Vol. 39), 163-199. San Diego: Academic Press.

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

Lamberts, K. (1994). Flexible tuning of similarity in exemplar-based categorization. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20, 1003-1021.

Lesgold, A., Rubinson, H., Feltovich, P., Glaser, R., Klopfer, D., & Wang, Y. (1988). Expertise in a complex skill: Diagnosing x-ray pictures. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), The Nature of Expertise (pp. 311-342). Hillsdale, NJ: Erlbaum.

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

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

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. Journal of Experimental Psychology: Human Learning and Memory, 7, 355-368.

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 (pp. 23-45). Chichester: Ellis Horwood.

Pazzani, M. J. (1991). Influence of prior knowledge on concept acquisition: Experimental and computational results. Journal of Experimental Psychology: Learning, Memory, and Cognition, 17, 416-432.

Ross, B. H., Perkins, S. J., & Tenpenny, P. L. (1990). Reminding-based category learning. Cognitive Psychology, 22, 460-492.

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

Taraban, R., & Palacios, J. M. (1993). Exemplar models and weighted cue models in category learning. In G. V. Nakamura, R. Taraban, & D. L. Medin (Eds.), Categorization by Humans and Machines (pp. 91-128). San Diego: Academic Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 1124-1131.

Ward, T. B. (1993). Processing biases, knowledge, and context in category formation. In G. V. Nakamura, R. Taraban, & D. L. Medin (Eds.), Categorization by Humans and Machines (pp. 257-282). San Diego: Academic Press.

Wattenmaker, W. D. (1995). Knowledge structures and linear separability: Integrating information in object and social categorization. Cognitive Psychology, 28, 274-328.

Wattenmaker, W. D., Dewey, G. I., Murphy, T. D., & Medin, D. L. (1986). Linear separability and concept learning: Context, relational properties, and concept naturalness. Cognitive Psychology, 18, 158-194.

Wisniewski, E. J. (1998). Property instantiation in conceptual combination. Memory & Cognition, 26, 1330-1347.

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

Table 1

Results and Predictions for Wattenmaker et al. (1986), Experiment 3.

Stimulus	Mean Errors	Integration	Mean Errors	Integration
	LS-Uninformed		LS-Informed	
A-111	0.6	0.13	0.7	0.14
A-110	3.3	0.29	4.1	0.29
A-101	2.7	0.37	4.4	0.38
A-011	2.9	0.34	3.9	0.33
B-000	1.1	0.13	1.7	0.16
B-001	1.7	0.29	3.0	0.30
B-010	3.9	0.37	6.5	0.43
B-100	3.5	0.34	4.4	0.35
Average	2.5	0.28	3.6	0.30
	NLS-Uninformed		NLS-Informed	
A-000	5.3	0.49	2.9	0.35
A-010	1.9	0.24	1.2	0.21
A-111	2.1	0.16	1.4	0.21
A-101	3.1	0.25	2.4	0.25
B-110	4.4	0.32	2.1	0.28
B-011	3.7	0.28	2.9	0.28
B-100	3.5	0.23	2.1	0.24
B-001	2.2	0.22	1.3	0.24
Average	3.3	0.28	2.0	0.26

Note. The results are stated in terms of mean errors during training. The model predictions are stated in terms of predicted error rate on a single trial.

Table 2

Results and Predictions for Barrett et al. (1993), Experiment 2.

Stimulus	Choice Proportion Animals	Integration Model	Stimulus	Choice Proportion Tools	Integration Model
11222	0.20	0.22	12122	0.17	0.18
22111	1.00	0.78	21211	0.83	0.83
1122*	0.40	0.18	1212*	0.17	0.18
112*2	0.00	0.04	121*2	0.00	0.03
11*22	0.20	0.25	1*122	0.17	0.14
2211*	0.60	0.78	2121*	0.83	0.82
221*1	1.00	0.96	212*1	0.83	0.97
22*11	0.80	0.85	2*211	1.00	0.86

Note. The results are stated in terms of mean proportion choosing the "Zibbot" category, which, during training, tended to have a 2 value on each dimension. On test stimuli, a * refers to a missing value.

Table 3

Summary of Murphy and Allopenna (1994)

Experiment 1

Condition	Learning Blocks	Error Rate: Integration
Arbitrary	8.6	.18
Meaningful	6.4	.18
Integrated	2.3	.12

Experiment 2

Condition	Error Rate: Results	Error Rate: Integration
Meaningful		
Frequent	.19	.13
Infrequent	.31	.35
Integrated		
Frequent	.08	.09
Infrequent	.08	.09

(table continues)

Experiment 3

Condition	Learning Blocks	Error Rate: Integration
Meaningful	5.2	.18
Domain	4.1	.18
Integrated	2.2	.15

Note. For all three experiments, the model predictions refer to predicted error rates on transfer trials. For Experiments 1 and 3, the results refer to average number of blocks to reach a criterion during training, and for Experiment 2 the results refer to error rates on transfer trials.

Appendix

Description of Modeling

The categorization model used in the analyses was a variant of the context model of classification (Medin & Schaffer, 1978). This model assumes that a classification decision on whether stimulus x belongs in category A or category B depends on the relative familiarity of x with respect to the members of the two categories. As shown in Equation A1, the probability of classifying x as an A rather than as a B increases with the familiarity of x with respect to category A (fam_A), and decreases with the familiarity of x with respect to category B (fam_B).

$$P(\text{classify x as A}) = \frac{fam_A(x)}{fam_A(x) + fam_B(x)} \qquad \text{(A1)}$$

The familiarity of a test stimulus is the total similarity of the stimulus to retrieved members of the category, as shown in Equations A2 and A3. To implement the integration process, in which categorization is influenced by both prior examples and observed category members, both sources of information are considered when evaluating familiarity. For example, in Equation A2, the familiarity of x depends on its similarity to the set of prior examples for category A, indicated by priorA, as well as its similarity to the retrieved observations of category A members, indicated by the set A. The k parameter varies from 0 to 1, and it refers to the relative influence of prior knowledge. When k is 0, integration of prior knowledge has no effect, and this categorization model is equivalent to the original context model.

$$fam_A(x) = (k) \sum_{pA \in priorA} sim(x, pA) + (1-k) \sum_{a \in A} sim(x, a) \qquad \text{(A2)}$$

$$fam_B(x) = (k) \sum_{pB \in priorB} sim(x, pB) + (1-k) \sum_{b \in B} sim(x, b) \qquad \text{(A3)}$$

The final aspect of the categorization model is the computation of similarity by the sim function, as described by Equation A4. In this multiplicative similarity rule, the similarity between stimuli x and y, each with n dimensions, is the product of their degree of

match along each of these dimensions. If $x_i = y_i$, that is, if x and y are the same on dimension i, then the match value, s_i, is 1. But if x and y differ on dimension i, then $s_i = w_i$, where w_i is some value between 0 and 1. In other words, when x and y agree on a dimension, the corresponding match value is 1, and when x and y differ, the match value is less than 1. The value of w_i indicates the degree of attention on that dimension, with lower values indicating greater attention.

$$sim(x,y) = \prod_{i=1}^{n} s_i \qquad \text{(A4)}$$

Implementation of Prior Examples

It is presumed that the sets priorA and priorB would include a large number of prior examples. For example, if A is a new category of birds, then many memories of birds might be retrieved. These prior examples would all have certain features, such as beaks and feathers, and vary on other dimensions, such as color. For the simulations run here, rather than representing this large number of prior examples and specifying all of the variable features, a small number of prior examples, with the characteristic features, was used. For the variable features, these prior examples used wild card values, which match any other value. So, one prior example with a beak and feathers might be represented in priorA, with a wild card value for color. Note that this assumption is a matter of convenience of simulation, and it is not critical to the integration model. In all of the simulations of the integration process reported here, equivalent performance may be obtained by increasing the number of prior examples and replacing the wild card values with specific values. Alternately, equivalent performance may be obtained by treating the wild card features as mismatches, and increasing the k parameter to compensate for the lower similarity of prior examples to observations.

It is also possible that some prior examples could be retrieved that are entirely dissimilar to what is observed in an experiment. For example, someone learning about a category of vehicles might retrieve prior examples of cars and space ships, and subsequently learn about new cars. The prior examples of space ships simply would not influence judgments; if the similarity between a new category member and a prior example is zero, then these prior examples would not affect the calculation of familiarity in Equation A2 or A3.

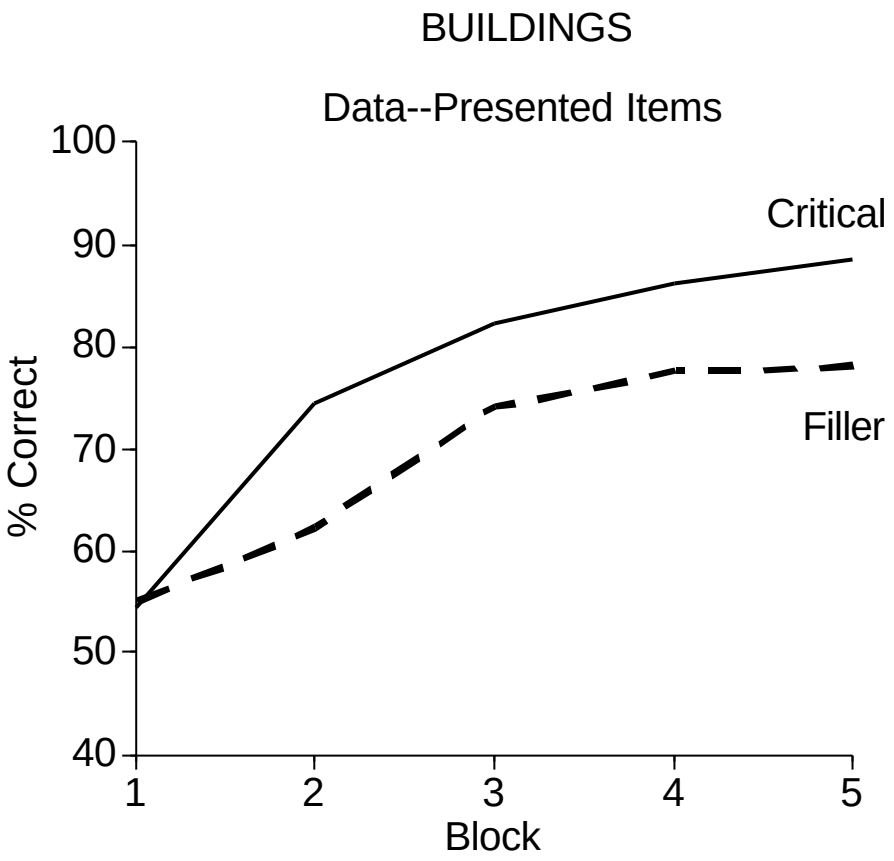

Figure 1. Results from Heit and Bott (1999), Experiment 2.

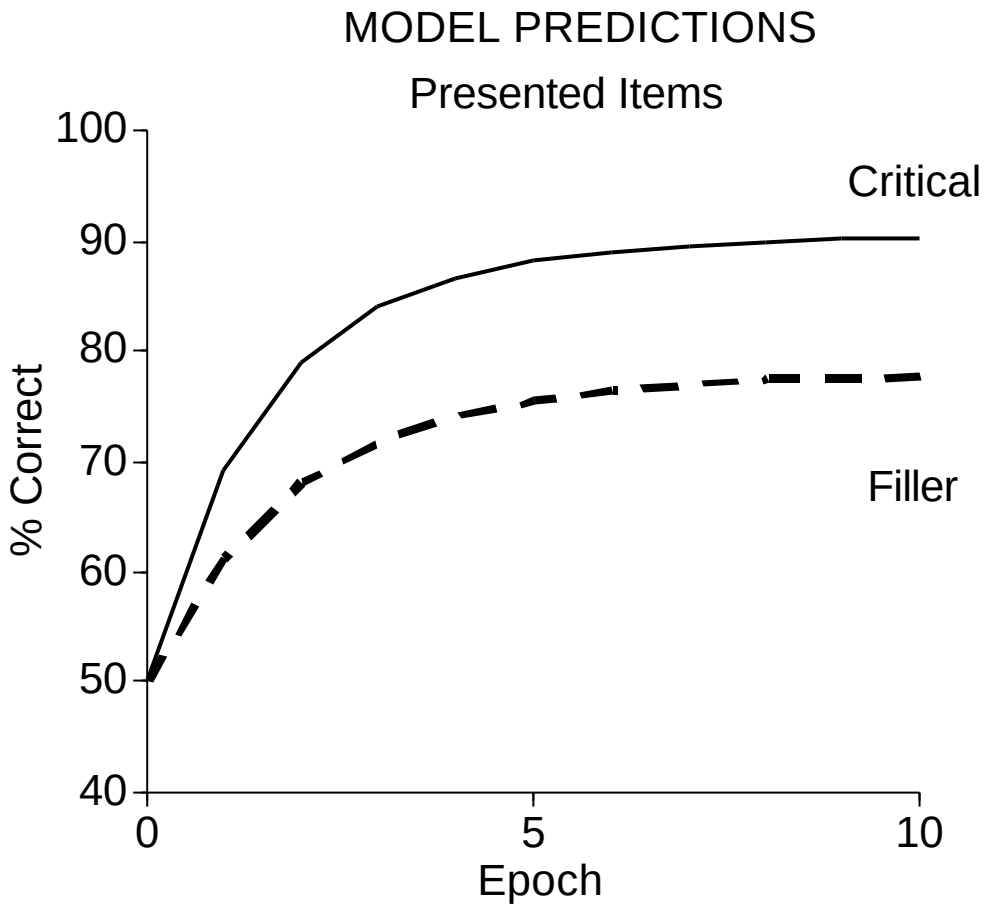

Figure 2. Predictions of the Baywatch model.