In D. H. Fisher, M. J. Pazzani, & P. Langley (Eds.), Concept Formation: Knowledge and Experience in Unsupervised Learning. San Mateo, CA: Morgan Kaufmann. (1991)

Harpoons and Long Sticks: The Interaction of Theory and Similarity in Rule Induction

EDWARD J. WISNIEWSKI
DOUGLAS L. MEDIN

1. Introduction

Concept learning is a fundamental aspect of intelligent behavior. Concepts let organisms relate new experiences to old ones, treat nonidentical experiences as equivalent, and make predictions about new experiences. All of these roles are crucial to an organism's survival. For example, consider the predictive role of a concept. By recognizing that a large, brown, furry creature seen from a distance is a bear, a person can predict that such a creature will be dangerous and avoid it. Given the significance of concepts, it is important to understand how intelligent systems form, update, and use them.

Understanding concept learning is a challenging problem. One can view concept formation as an *unsupervised* task of partitioning n items into a set of categories (i.e., equivalence classes) and forming a concept for each category (e.g., Anderson & Matessa, this volume; Fisher, 1987). In this context, a concept is a classification rule that divides items into those that satisfy the rule (i.e., those belonging to the category) and those that do not. However, there are an exponential number of ways in which n entities can be partitioned into categories; for any n of reasonable size, the learner cannot possibly consider all of them (Anderson & Matessa, this volume).

Even in more constrained, *supervised* tasks (Fisher & Pazzani, Chapter 1, this volume), in which the partition is known and when the learner is presented with preclassified entities, there may be many possible classification rules that apply to each category. Therefore, machine learning has been greatly concerned with discovering useful biases or constraints on concept learning (e.g., Mitchell, 1980; Utgoff, 1986). Psychological studies of concept learning also have attempted to determine the biases that make some categorization tasks easy and natural and others difficult and unnatural (e.g., Medin, 1983; Medin & Schwanenflugal, 1981).

More generally, cognitive psychologists and researchers in machine learning have recently discovered a common agenda with respect to understanding concept learning. Cognitive psychologists have become interested in treating machine learning programs as candidates for psychological process models, and machine learning researchers have turned to psychological research to identify useful constraints (e.g., Fisher, 1988; Fisher & Langley, 1990; Medin, Wattenmaker, & Michalski, 1987; Pazzani, Dyer, & Flowers, 1987; Pazzani & Schulenburg, 1989). In both disciplines, a prominent approach to concept learning relies on empirical or data-driven learning (e.g., Anderson & Matessa, this volume; Dietterich, London, Clarkson, & Dromey, 1982; Fisher, 1987; Hintzman, 1986; Medin & Schaffer, 1978; Michalski, 1983a, 1983b; Mitchell, 1982; Nosofsky, 1986; Quinlan, 1983, 1986; Smith & Medin, 1981). Typically, this strategy involves acquiring concepts based on patterns of similarities and differences that are observed across a number of training items. However, researchers in psychology have increasingly emphasized the limitations of similarity as the basis for concept learning.

A major problem is that empirical learning methods can be misled by irrelevant information (Schank, Collins, & Hunter, 1986). For example, suppose that by coincidence, all the apartment dogs that a learner has been exposed to were brown. Many empirical models would include this irrelevant feature in the concept "apartment dog". This problem is accentuated if there are not a significant number of examples available, or if examples are "costly" to obtain. Consider one example (Scott, 1987) of a system that places a can of Coca Cola in the freezer to cool and returns hours later to discover that it has shattered. There are many features (and interactions among these features) associated with this situation (e.g., the shape, color, composition, and contents of the can, the shape and color of the freezer). To empirically learn why the can shattered could require that the system observe many examples,

noting which features vary and which remain constant. Such a scenario would be impractical and expensive.

In response, many researchers argue that human concepts are organized around people's theories about objects and events in their world (e.g., Carey, 1985; Keil, 1981; Murphy & Medin, 1985; Rips, 1989). In other words, learning is *theory driven*. For example, a theory-driven system could learn why the Coca Cola can shattered from a single example, by reasoning from its prior knowledge that water expands when it freezes, that Coca Cola is composed almost entirely of water, and so on. It also might generalize this explanation to other beverages, containers, and situations involving low temperatures. Similar motivations for theory-driven processing have led machine learning researchers to focus on *explanation-based learning* (Mitchell, Keller, & Kedar-Cabelli, 1986; DeJong & Mooney, 1986; Ellman, 1989; Mooney, this volume). Typically, an explanation-based system uses its background knowledge (in the form of a theory) to explain or prove why a training example is a member of a given category. It then generalizes the explanation so that it will apply to future examples. In both the human and machine case, theories provide biases or constraints on learning.

However, theory-based approaches also have their problems. Typically, explanation-based learning systems learn by restructuring their existing knowledge. If a learner's domain theory is incorrect, incomplete, or inconsistent, then it may incorrectly explain or fail to explain a given phenomenon (Mitchell, Keller, & Kedar-Cabelli, 1986; Rajamoney, 1986). In addition, many concepts have important non-explanatory components that are often conventional in nature (Ahn & Brewer, 1988; Mooney & Ourston, 1989). For example, features like `has a white color` and `has a circular shape` typically would not enter into an explanation for why a light bulb gives off illumination. An explanation-based learning system would fail to incorporate these features into its `light bulb` concept, but they are nonetheless important for identifying light bulbs.

As implied above, some of the strengths of each approach complement the weaknesses of the other. As a result, cognitive psychologists have argued for the integration of empirical and theory-driven learning (Medin & Ortony, 1989; Wattenmaker, Nakamura, & Medin, 1987) and researchers in machine learning have developed a number of systems that combine both empirical and explanation-based learning (e.g., Flann &

Dietterich, 1989; Kedar-Cabelli, 1985; Lebowitz, 1986a, 1986b; Mooney & Ourston, 1989; Pazzani, 1987, 1988; Rajamoney, 1986; Shavlik & Towell, 1989; Wisniewski, Winston, Smith, & Kleyn, 1987).

Despite the concern with combining these two learning paradigms, most current approaches do not tightly couple their interaction. In this chapter we argue that, at least in some domains, this loose coupling is inadequate. We begin our analysis by describing a standard concept learning task (rule induction) found in many psychological experiments. Next, we present several recent studies that have examined the relation between theories and data in rule formation. We then explore the ways that researchers have combined explanation-based and empirical learning. The plausibility of these approaches can be evaluated in light of our own experimental findings, which expose the close interaction between theory and similarity in concept learning. We conclude the chapter by briefly sketching a model of learning in which theory and data are more tightly coupled.

2. Rule Induction Paradigms

In a typical rule induction task, the experimenter selects a rule or concept, and participants must learn the rule based on feedback that they receive from classifying examples (e.g., Bruner, Goodnow, & Austin, 1956; Haygood & Bourne, 1965). Furthermore, participants may be told what types of rules are involved (e.g., conjunctive rules, single feature rules), so that learning involves a selection among a small set of rules. In the tasks that we will describe, however, the experimenter does not select a particular rule for the subject to learn. Rather, there are many possible rules that describe the categories and interest lies in which rules people find natural to form. The rules that people create should reveal constraints on induction that can be used to evaluate alternative models of rule formation.

2.1 Simultaneous and Sequential Induction

We will describe studies that involve two types of rule induction tasks, which closely correspond to nonincremental and incremental models discussed by Fisher and Pazzani (Chapter 1, this volume). In the *simultaneous* rule induction task, participants determine a single rule by examining a number of preclassified items that are presented at the same

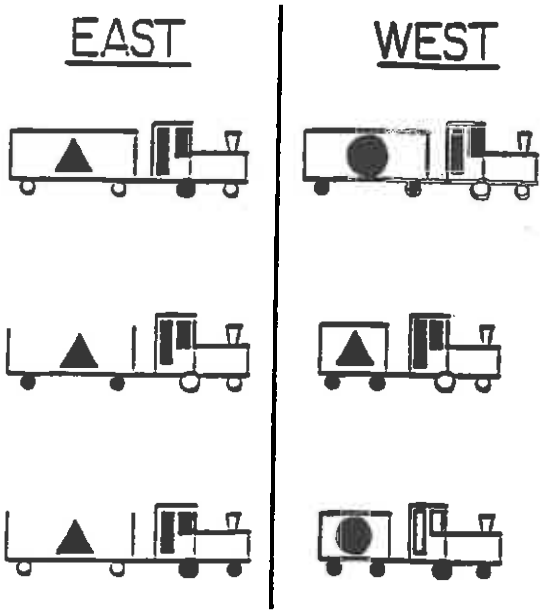

Figure 1. Stimuli used in concept induction tasks by Medin, Wattenmaker, and Michalski (1987).

time. In the *sequential* rule induction task, training items are presented sequentially rather than simultaneously and participants determine a rule after seeing each item. Furthermore, the items are not preclassified. Instead, participants determine a rule that can be used to classify items into a category, and then receive feedback on their classification. Thus, rules may change over the course of the task.

Figure 1 shows stimuli used in one simple rule induction task, taken from Medin, Wattenmaker, and Michalski (1987). Here, a set of trains have been preclassified as `Eastbound` or `Westbound`. There are many possible rules that apply to these categories. Rules for the `Eastbound` trains include "long car and triangle load", "white car wheels or open car top", "long car and not circular load", "open car top or engine with one white wheel", and so on.

One reason that the rule induction task is of interest is that alternative learning systems will induce different rules and one can compare these rules to those that people develop. For example, Quinlan's (1986) ID3 would favor simple disjunctive rules, such as `white wheels or open car`, for the examples shown in Figure 1; in contrast, Michalski's (1983a,

1983b) INDUCE would favor the rule long car and triangle load, though for other category structures INDUCE might yield a disjunctive rule. Medin et al. (1987) found that people were far more likely to develop conjunctive rules than disjunctive rules for the trains shown in Figure 1. The major difference between the rules formed by INDUCE and the participants was that people were somewhat more likely to begin with a simple rule that was complete (i.e., applied to all positive examples) but inconsistent (i.e., also applied to nonmembers). They would then refine the rule to make it both complete and consistent (i.e., apply to all members and only to members of the category). For example, people might start with a rule like Eastbound trains have triangle loads, then notice the counterexample and amend their rule to Eastbound trains have a triangle load and not a short car (see Figure 1).

2.2 Using Theories to Determine the Feature Space

In both cognitive psychology and machine learning, researchers often investigate learning by providing intelligent systems with a space of well-defined features that describe one or more training items. Learning is viewed as selecting an appropriate combination of features from this space. This view is especially common in traditional empirical learning systems (see Fisher & Pazzani, Chapter 1, this volume). However, a crucial problem in learning involves deciding the units of analysis or constituents upon which to invoke learning (Medin, Wattenmaker, & Michalski, 1987). That is, learning involves not only selecting features from a feature space but also determining that feature space. This problem typically is solved for the learning system by the programmer, who presents the system with predefined, well-specified, unambiguous constituents. How does one determine the feature space upon which learning operates?

Explanation-based learning provides a starting point for exploring how theories determine the feature space (Mooney, this volume). Typically, an explanation-based system is given a training item, described as a set of features (e.g., a particular item with features such as weight of item is light and item has a handle) and a functional specification (e.g., one can drink from the item). Using a theory, it deductively proves that or explains why the particular item meets this functional specification. The explanation is actually a tree whose root is the functional specification and whose leaves are a subset of the features of the

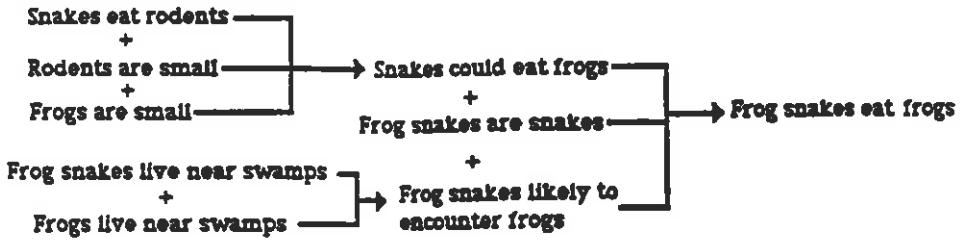

Figure 2. A plausible theory for why frog snakes eat frogs.

training example. The intermediate nodes of the tree are additional features that explanation-based learning has deduced about the item.

As we will suggest in this chapter, people's theories and how they use them to determine features is quite different from the approach taken in explanation-based learning. For one thing, an explanation-based system typically views a theory as a well-defined set of rules that are used to construct a deductive explanation (e.g., Mitchell et al., 1986; Yoo & Fisher, this volume). In contrast, people's everyday theories contain a mixture of fact and fiction. A person's theories may combine scientific principles, stereotypes, and informal observations of experiences. Furthermore, people use these theories to construct plausible rather than deductive explanations (Collins & Michalski, 1989). For example, suppose you read about a novel reptile called a *frog snake*, which was found near swamps in South America. You might reasonably conclude that frog snakes eat frogs. Figure 2 illustrates a simple theory and explanation for why this might be the case. Some of the reasoning may follow from informal observations. For example, you may believe that snakes generally eat rodents simply because you saw a TV program in which a snake ate a mouse. Thus, the explanation is plausible rather than deductive, and it could well turn out that frog snakes do not eat frogs. Rather, they might be called frog snakes because they croak like a frog or because they have bulging, frog-like eyes.

How might one study theories in categorization tasks? One approach involves using complex rather than simple stimuli. The relatively simple train stimuli of Figure 1 have the virtue that they can be used to directly compare the rules formed by induction programs with those formed by people. The constituents (features) of the trains that a programmer would present to induction programs are probably the same ones that people would consider in forming their rules. The stimuli have relatively few, unambiguous constituents (e.g., `wheels`, `color of wheels`, and `type of load`).

However, these virtues are also limitations: a small, unambiguous feature set is often an unrealistic presumption about a domain. Additionally, people are very unlikely to have elaborate theories about the behavior and appearance of `Eastbound` and `Westbound` trains. Therefore, we have shifted our attention to more complex stimuli that are likely to involve people's theories and to highlight the problem of determining a feature space.

In particular, the category items for our experiments are children's drawings of people, shown in Figures 3, 4, and 5. The drawings, taken from Koppitz (1984) and Harris (1963), were produced by children who were administered the "draw-a-person test". In this task, children are instructed to draw "one whole person". This test is one tool used in psychodiagnosis and IQ assessment, and it is a fairly reliable shorthand indicator of emotional problems and of intelligence in young children (Goodenough & Harris, 1950; Harris, 1963; Koppitz, 1984).

Given these more complex stimuli, one simple way of teasing out theoretical preconceptions is by providing subjects with meaningful labels for the categories (e.g., Adelman, 1981; Muchinsky & Dudycha, 1974; Wattenmaker, Dewey, Murphy, & Medin, 1986; Wisniewski & Medin, 1991; Wright & Murphy, 1984). The purpose of providing such labels is to activate theories or prior expectations that may guide rule induction, in much the same way that functional specifications do in explanation-based learning. We can contrast *theory-guided* rule induction to the situation described above, in which the category labels have little meaningful content and therefore theories are less likely to be activated.[1] We

1. Prior expectations also could be triggered by information in the training instances. For example, even without the label done by creative children, a person might believe that a particularly realistic, detailed drawing was done by a creative child. As a result, that person might hypothesize that drawings of a category were done by such children. However, this was seldom the case.

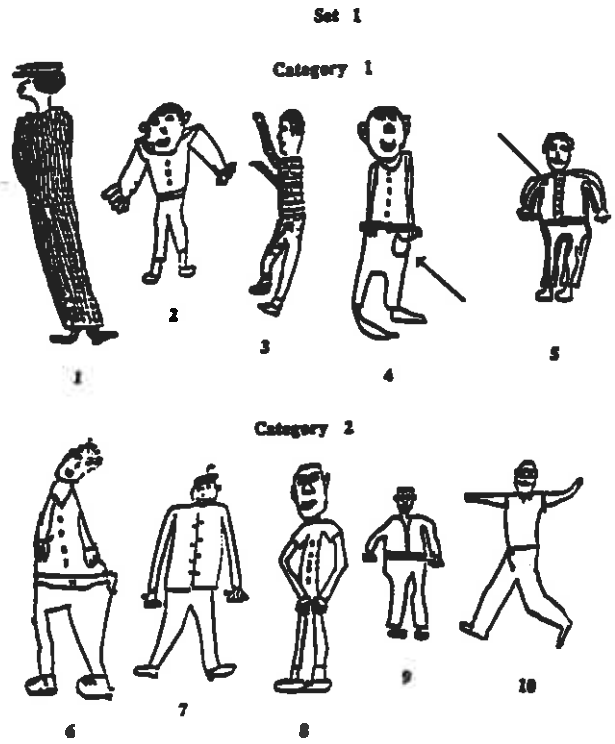

Figure 3. Drawings by "high" and "low" IQ children in Study 1, and by "farm/city" and "creative/noncreative" children in Study 2.

will call this situation *empirical* rule induction. The latter is more typical of categorization studies in psychology (e.g., Hintzman, 1986; Medin, Wattenmaker, & Michalski, 1987; Nosofsky, 1986) and a number of machine learning programs (Fisher, 1987; Michalski, 1983b; Mitchell, 1982; Quinlan, 1986; Winston, 1975).

2.3 Supervised Versus Unsupervised Learning

The rule induction tasks that we have described are supervised. That is, the experimenter indicates to the participant the category in which an item belongs. In contrast, *unsupervised* tasks do not involve explicit feedback from an experimenter (or programmer). For example, conceptual clustering systems (e.g., Anderson & Matessa, this volume; Fisher, 1987; Hanson & Bauer, 1989) partition items into categories based on some evaluation function, forming a concept or rule for each category. Learning is unsupervised in the sense that a programmer does not give the system feedback about whether or not an item belongs in a category.

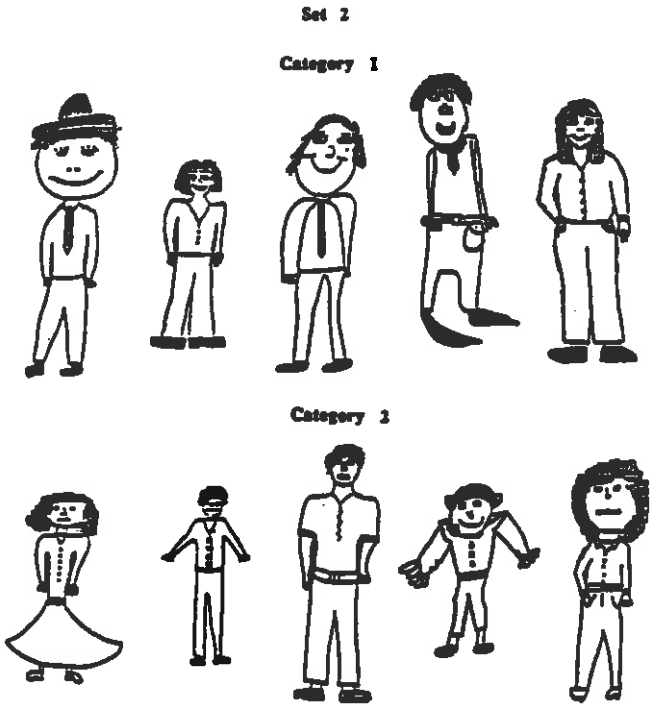

Figure 4. Drawings by "creative" and "noncreative" children.

In general, how important is the distinction between supervised and unsupervised learning? Langley (1987) suggests that unsupervised learning characterizes much of human learning. He notes that before they can understand language, children form useful concepts primarily by direct interaction with their environment, rather than by advice given to them by other humans. On the other hand, it is clear that in many domains, human learning is supervised. Learning medical diagnosis, nuclear reactor monitoring, mathematical problem solving, and guitar playing are just a few of the domains in which people are explicitly informed of the categories in which items belong.

We believe that the distinction between supervised and unsupervised learning is really addressing a deeper issue: the nature of feedback and credit assignment. All systems eventually must use their concepts; otherwise, there would be little point in learning them. Importantly, some type of success or failure (i.e., feedback) will be associated with that use. Thus, it seems reasonable that intelligent systems should take advantage of feedback and assign credit to concepts that lead to success

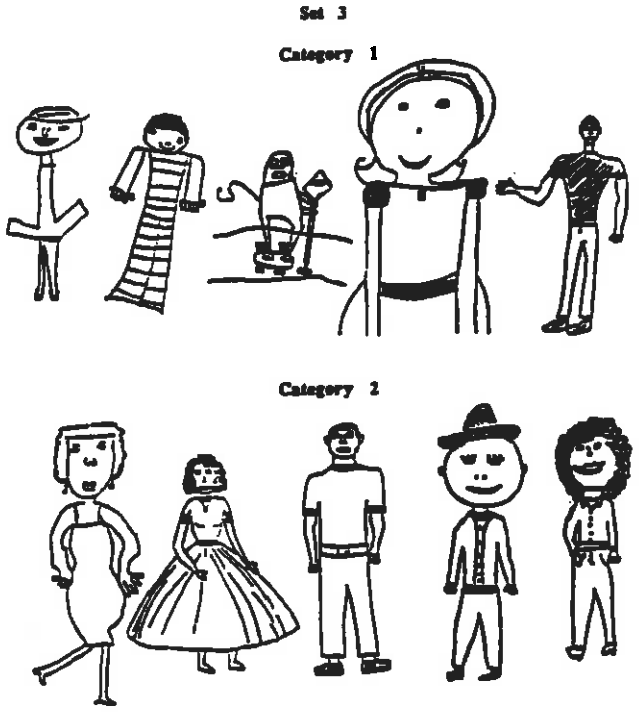

Figure 5. Drawings by "well-adjusted" and "emotionally-disturbed" children.

and blame to concepts that result in failure. In any event, we believe that the distinction between supervised and unsupervised learning is not crucial to the studies that we will describe. In these studies, even with explicit feedback, credit assignment was not at all straightforward. As we describe later, our stimuli were more complex than those typically found in psychological studies and in computational simulations. Determining the reason a group of items belonged to the same category was not an easy task. This was especially true since the validity of a person's hypothesized rule could vary with its level of abstraction. For example, consider a person who believes that a drawing is `detailed` because it has `hair`, `eyes`, and `shoes`. At a high level of abstraction, the person might correctly assume that the drawing belongs to a particular category because it is `detailed`. At the same time, the person's more specific reasons for why the drawing is detailed (i.e., `hairs`, `eyes`, and `shoes`) might be incorrect. (The drawing might be `detailed` because it has `curly hair`, `eyebrows`, and `shoelaces`.) In addition, a goal of our second study was to examine people's rule modification strategies. In this case, explicit negative feedback was desirable.

3. Study 1: Simultaneous Rule Induction

The first study explored the effects of theoretical knowledge, empirical knowledge, and their interaction in a simultaneous rule induction task. In this task, participants determined a classification rule by examining a number of the members of a category that were presented at the same time. The categories consisted of the three sets of children's drawings shown in Figures 3, 4, and 5. Two groups of participants, the THEORY 1 and THEORY 2 groups, learned about pairs of categories with meaningful labels (indicated in Figures 3, 4, and 5). In contrast, the STANDARD group learned the identical pairs of categories, but were given irrelevant labels. The basic task was straightforward. Participants examined each set of two categories and wrote down a rule that distinguished the members of one category from those of the other. In addition, the THEORY 1 and THEORY 2 groups differed in that the labeling of their categories was counterbalanced. For example, in Figure 3, the THEORY 1 group was told that Category 1 was drawn by `low IQ children` and that Category 2 was drawn by `high IQ children`. For the THEORY 2 group, this labeling was reversed.

3.1 Goals of the Study

This study had two purposes. First, it should provide a simple demonstration of whether people use integrated or nonintegrated learning. Because both empirical and theoretical knowledge are available, it should tell whether learning is primarily empirical, theory guided, or some combination of the two. A number of findings are consistent with each of these possibilities. For example, if the THEORY groups use only empirical learning and ignore theoretical knowledge (activated by the meaningful labels), then their rules should be similar to those of the STANDARD group. This would occur because all groups learn about *identical* categories and therefore are provided with the *same* empirical knowledge. There are a number of rules that can be found empirically (i.e., without recourse to theoretical knowledge) that distinguish the categories in each set. For example, in Figure 4, the rule "smiling and hands at side" distinguishes the members of Category 1 from those of Category 2.

On the other hand, if people use only their theories and ignore empirical evidence, then the THEORY 1 and THEORY 2 groups should produce similar rules for identically labeled but different categories. For

example, the rules of the THEORY 1 group for Category 1 in Figure 3 should be very similar to those of the THEORY 2 group for Category 2. This would occur because both groups were told that these different categories were drawn by low IQ children. In this case, *identical* theoretical knowledge should have been activated in these groups, even though it is present in the context of *different* empirical knowledge.

Although the two situations described above are logical possibilities, they represent rather extreme views of how people utilize theoretical and empirical knowledge. Each view suggests that with both sources of knowledge available, people ignore one source and only use the other. More than likely, however, people use some combination of theoretical and empirical knowledge. There would be good evidence for such integrated learning if the rules of the STANDARD and THEORY groups were different and if the rules for the THEORY 1 and THEORY 2 groups were different for identical but differently labeled categories.

A second purpose of the study was more exploratory in nature. We were interested in the effects of theoretical expectations on rules. Specifically, how might the rules of people with expectations differ from those of people without such expectations? To explore this issue, we compared the rules of the THEORY groups to those of the STANDARD group.

3.2 Method

The subjects were 40 undergraduates (male and female), attending the University of Michigan, who received course credit for participating in the study. A participant was assigned randomly to a THEORY group or the STANDARD group. Each participant saw three sets of children's drawings, with each set divided into two equal-sized categories of five drawings (as shown in Figures 3, 4, and 5).

Participants in the study were instructed to write down a rule for each set that someone else could use to accurately place the drawings in their respective categories. They also were told that someone else should be able to use the rule to decide whether a new (not yet seen) drawing belonged to one or the other category. For the STANDARD group, the two categories in each set were simply labeled Category 1 and Category 2. For the THEORY groups, the two categories in each set were given meaningful labels, corresponding to the functional specifications typically given to explanation-based learning systems. We used different meaningful labels to explore a variety of theories that might

be activated. Specifically, the two categories in the first set (Figure 3) were labeled drawn by children with high IQ/low IQ, those in the second (Figure 4) were labeled drawn by creative/noncreative children, and those in the third set (Figure 5) were labeled as drawn by emotionally disturbed/well-adjusted children. As noted, the labeling of categories was counterbalanced in the THEORY groups. That is, for the THEORY 1 group, a category was given one of the two meaningful labels, and for the THEORY 2 group, the same category was given the other meaningful label.

3.3 Results

To explore possible learning strategies (i.e., empirical only, theory only, or some combination), we compared the rules of the STANDARD group to those of THEORY groups and the rules of the THEORY 1 group to those of the THEORY 2 group. In general, there were differences between the THEORY groups and the STANDARD group and between the two THEORY groups. These differences suggest that people use some combination of theoretical and empirical knowledge to formulate rules. We describe these results below.

3.3.1 EFFECTS OF THEORETICAL KNOWLEDGE ON RULE LEARNING

To examine the impact of theoretical knowledge, we compared the similarity *between* the rules of the THEORY groups and those of the STANDARD group to the similarity of the rules *within* the STANDARD group for the same category. If theoretical knowledge affects rule formation, then on the average, a rule from the THEORY group and one from the STANDARD group should be more different than two rules from the STANDARD group. A rule from the THEORY group should be based on both theoretical and empirical knowledge, whereas a rule from the STANDARD group will be based on only empirical knowledge. In contrast, any two rules from the STANDARD group will be based on the same empirical knowledge (if they apply to the same category).

We measured the similarity of rules by calculating their feature overlap. More specifically, for each rule in the THEORY groups, we calculated the percentage of rules in the STANDARD group that matched at least one constituent in the THEORY-group rule. We then averaged these percentages. Across categories, the average percentage overlap between

the rules of the THEORY and STANDARD groups was 15%. To measure the similarity of the rules in the STANDARD group, we randomly divided the rules of a category into two equal-sized groups. For each rule in one group, we calculated its average percentage overlap with the rules of the other group. Across categories, the average percentage overlap of rules within the STANDARD group was 26%. The fact that the similarity of the rules of the THEORY groups and STANDARD group (15%) was less than the similarity within the STANDARD group (26%) suggests that theoretical knowledge affected the rules of the THEORY groups.

One curious issue is why the rules of the STANDARD group differed from each other so much for the same category. Although the features of these rules overlapped more with each other than with the rules of the THEORY group, the overlap was only 25%. There are probably several reasons for this result. First, the stimuli are relatively complex and there was a wide range of features that people could use in their rules. Second, as detailed later, very few features were complete and consistent, in that they applied to all members of a category and to none of the contrast category. In a sense then, many features may have been "equally good" for determining category membership and participants were not biased to use any particular one in forming their rules.

3.3.2 EFFECTS OF EMPIRICAL KNOWLEDGE ON RULE INDUCTION

To study the effect of empirical knowledge, we examined the rules of the THEORY 1 and THEORY 2 groups for different categories that were identically labeled. In these cases, identical theoretical knowledge but different empirical knowledge should be available to these groups. In general, the rules of the groups were very different, suggesting that people were not just using theoretical knowledge. In fact, the average percentage overlap between the rules of the THEORY 1 and THEORY 2 groups for identically labeled categories was only 2.4% using the measure described above.

To get some appreciation of these differences, consider the set of categories shown in Figure 5. For Category 1, the THEORY 1 group mentioned `lack of detail` or `simplicity` in four of their eight rules, when the category was labeled `drawn by emotionally disturbed children`. In contrast, the THEORY 2 group failed to mention `lack of detail` in any of their rules for the identically labeled Category 2. Instead, the THEORY 2 group mentioned that the drawings of Category 2 were

accurate or precise (three of nine rules) or that they depicted normal, conventional, or regular looking people (five of nine rules). None of these characteristics was mentioned in any of the rules of the THEORY 1 group for the identically labeled Category 1. In short, empirical knowledge also appeared to affect the rules of the THEORY groups. Based on these analyses, it appears that people in the THEORY groups used some combination of theoretical and empirical knowledge in formulating rules. We now take a closer look at how people's theories affected rule formation.

3.3.3 THEORY GROUP VERSUS STANDARD GROUP RULES

There were some striking differences between the rules of the THEORY groups and those of the STANDARD group. First, we noticed that the two groups differed in the kinds of features that they used in their rules. We classified the rules into three major types, based on the types of features that they included. *Concrete* rules consisted of simple features that were easily observable in the drawings (e.g., eyebrows, white shoes, buttons, ears, striped clothes). *Abstract* rules consisted of features that were more complex, higher level, or less perceptual (e.g., normal, bizarre, detailed, well-proportioned, hastily drawn). *Linked* rules consisted of both concrete and abstract features (e.g., cross-eyes, abnormalities, proportion).

Two examples of each rule type for categories 1 and 2 of Figure 3 illustrate the differences:

- *Concrete*: "buttons or stripes on their shirts and dark, thick hair" (for Category 1) and "facing forward and showing ears and not showing teeth" (for Category 2);
- *Abstract*: "relaxed and free flowing" (for Category 1) and "look more normal" (for Category 2);
- *Linked*: "added more details such as teeth, extensive shading, and drawing the body underneath the clothes" (for Category 1) and "limbs proportional to rest of the body, some drawings demonstrate dimensional representation (e.g., one leg crossing over the other), more detailed dress (e.g., shoelaces)" (for Category 2).

Table 1 presents the percentages of each type for the two groups. In using linked rules, people often listed several different features as exam-

Table 1. Frequency of rule types among the STANDARD and THEORY groups.

	STANDARD GROUP	THEORY GROUP
% CONCRETE RULES	81	35
% ABSTRACT RULES	16	37
% LINKED RULES	3	28

ples of a more abstract feature. The linked rules provide evidence that the THEORY group *operationalized* abstract information (e.g., Mostow, 1983) by linking it to concrete features in the data.

If this account of operationalization is correct, then one might ask why the THEORY groups produced fewer linked rules than abstract rules. One reason is that there may have been many different ways to operationalize an abstract feature among the examples. People may have written their rules at the level of description that was common to all examples. A similar strategy is used by some systems that combine explanation-based and empirical learning, such as Flann and Dietterich's (1989) IOE and Yoo and Fisher's (this volume) EXOR system. Second, the rules may have been less detailed because of such subtle factors as the amount of space provided for writing and the time spent writing. In a previous study involving meaningful category labels (done with Glenn Nakamura), almost all the rules produced were linked. In that study, participants had more space and time to write down rules.

A second observed difference was that the rules of the STANDARD group generally were more syntactically complex than those of the THEORY group. To measure syntactic complexity, we counted the number of properties and logical operators (conjunction, disjunction, or negation) used in each rule. For the STANDARD group, 59% of the rules contained more than two features, compared to 44% of the rules for the THEORY group. The STANDARD group used an average of 2.42 logical operators per rule, compared to 1.60 for the THEORY group.

To summarize, there were some clear differences between the kinds of features occurring in rules of the THEORY groups and those of the STANDARD group. The latter primarily used concrete rules, seldom used abstract rules, and very rarely used linked rules. In contrast, the THEORY groups used many fewer concrete rules, and many more abstract

and linked rules, than did the STANDARD group. The groups also differed in terms of the syntactic complexity of their rules. In general, the STANDARD group used rules that contained more features and more disjunctions, conjunctions, and negations of properties.

3.4 Discussion

Our results suggest that the THEORY group combined both theoretical and empirical knowledge in constructing their rules. In particular, it appears that theoretical knowledge provides certain biases to the learner. One way that theoretical knowledge strongly affects rule formation is by activating a space of abstract features and hypotheses. People are then biased to search for evidence in the data that supports these features and hypotheses. Informal examination of the linked rules suggests that this was the case. For example, consider the following linked rule:

> High IQ children are capable of drawing people from a profile angle; their drawings show detail such as pockets and collars in their figures.

Figure 6 illustrates that the label, drawn by high IQ children, may have activated the hypothesis that intelligence is required to draw well and, therefore, more intelligent children will draw better. This hypothesis, in turn, might suggest that better drawings will capture abstract features such as details (e.g., pockets and collars) and difficult aspects of drawing people (e.g., profile angles).

By activating high-level, abstract features, theories also may bias people to treat lower-level features as equivalent, when in other contexts they would be considered different. In other words, theories allow people to generalize across different features. For example, in a neutral context, the features pockets and collars appear quite different. However, as illustrated in Figure 6, the theory-activated expectation that drawings will show detail lets one generalize across these features and consider them similar.

Most empirical learning systems do not generalize across features and do not represent features at multiple levels of abstraction, but there are exceptions. Some inductive systems use various heuristics to treat features equivalently. For example, INDUCE (Michalski, 1983a) uses an IS-A hierarchy to transform the features Chicago, De Kalb, and Peoria into the feature Cities in Illinois. However, note that such a hier-

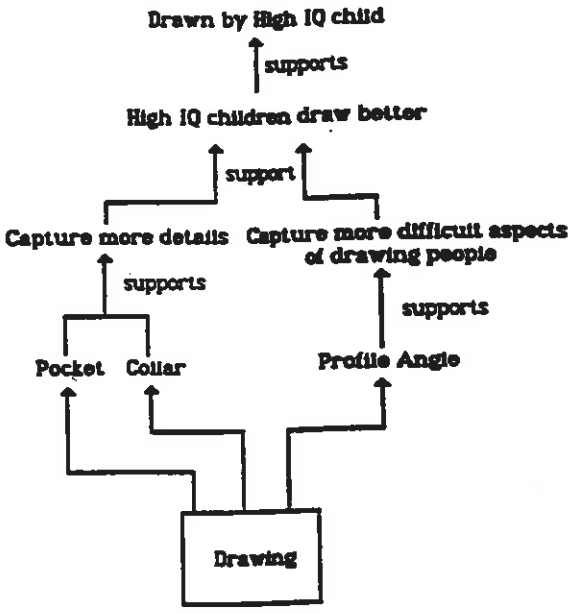

Figure 6. Plausible explanation for a drawing by a high IQ child.

archy is built into the system. In our studies, people probably created such hierarchies dynamically, through an interaction of empirical and theoretical knowledge. It seems unlikely that our subjects (typical undergraduates) had prestored, hierarchical knowledge about drawings of people done by emotionally disturbed children.

Whether or not a system generalizes across features has important effects on how it carries out induction. To take one simple example, consider the feature descriptions of four children's drawings shown in Figure 7. These descriptions correspond to actual drawings done by well-adjusted and emotionally disturbed children.[2] Consider two induction systems whose task is to cluster the descriptions into categories. The first system groups descriptions based on a measure of surface similarity like *category utility* (Gluck & Corter, 1985; Fisher & Pazzani, Chapter 1, this volume) and does not generalize across features. The second system groups descriptions based on theoretical expectations

2. In children's drawings, one indicator of emotional problems is the omission of body parts (Koppitz, 1984). In the drawings done by children with emotional problems, note that one description includes the features hands missing and nose missing, whereas the other description includes the features feet missing and neck missing.

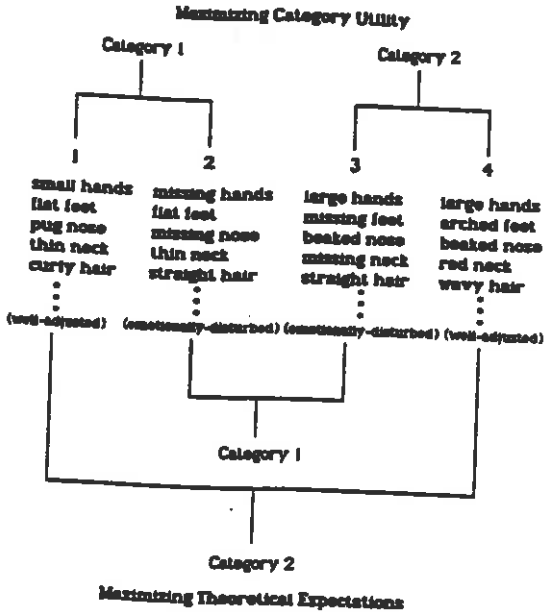

Figure 7. Alternative classifications of drawings by well-adjusted and by emotionally disturbed children.

about `well-adjusted` and `emotionally disturbed` children and generalizes across features. These systems will cluster the descriptions in different ways, as shown in Figure 7. Specifically, by grouping drawings 1 and 2 together and drawings 3 and 4 together, the first system maximizes surface similarity in terms of category utility. In contrast, the second system will group descriptions 2 and 3 together and 1 and 4 together. Such a system can treat the features `hands missing`, `nose missing`, `feet missing`, and `neck missing` as equivalent, because they are all examples of the same high-level feature, `missing body parts`. This feature, in turn, is an example of `deviation from normal`, which in turn provides evidence for the category `emotionally disturbed`. On the other hand, the second system will group descriptions 1 and 4 together, as these examples do not deviate from the normal, thus implying that both were done by `well-adjusted` children. Interestingly, at the concrete feature level, these descriptions have few features in common, yet were placed in the same category.[3]

3. Of course, one could include `missing body part` as a feature in the drawings by the `emotionally disturbed` children, and then highly weight this feature during categorization. In this case, a model that maximizes category validity could derive the alternative, theory-based grouping, but there would be no clear

In contrast to the THEORY groups, the STANDARD group did not operate with the biases described above. Its members seldom used abstract rules or linked rules involving operationalization, suggesting that abstract features and hypotheses were absent. They did not generalize across features. Instead, this group appeared to focus on finding any set of simple, observable features that could discriminate the categories.

The STANDARD group constructed rules that were considerably more complex than those typically found in other studies of empirical learning. For example, in the study involving the train stimuli shown in Figure 1, Medin et al. (1987) found that 75% of the rules were based on a single property, a conjunction of two properties, or a disjunction of two properties. Only 41% of the rules of the STANDARD groups could be classified as one of these three types. Instead, many rules involved complicated combinations of conjunctions and disjunctions, with three or more properties. Most likely, the rules of the STANDARD group were more syntactically complex because they relied on concrete features that were neither consistent nor complete. A feature is consistent if it is present in (at least) some members of a category but absent in all members of the contrasting category. A feature is complete if it is present in all members of the category. Simple rules, like the three types noted in the Medin et al. study above, typically are constructed out of features that have these characteristics. A correct rule can be based on a single feature if that feature is both complete and consistent. It can be based on conjunction of two features if each feature is complete and their conjunction is consistent. Finally, a rule can be based on a disjunction of two features if each feature is consistent and their disjunction is complete. Looking at the train stimuli shown in Figure 1, one can see that there are many features that have one or both of these characteristics. Not surprisingly then, a large majority of rules in the Medin et al. (1987) study were simple rules.

For the categories used in our drawings, however, there were few features with one or both of these characteristics. Specifically, for each pair of categories, we examined the completeness and consistency of 40 concrete features that were selected arbitrarily from the rules given by subjects. For the three category pairs, none of the features was

justification for including the feature and weighting it highly. Our approach includes the feature because different, low-level features are viewed as similar, high-level features. Their importance comes from prior expectations about the category.

complete and consistent, 32% were consistent, and 11% were complete. In addition, for those features that were consistent, almost all of them (87%) were true of only one or two (out of five) category members. Also, the most syntactically complex rules were those for the categories shown in Figure 3, which had fewer complete or consistent features than other category sets. Specifically, none of the features was complete for these categories, and all those that were consistent were true of only one or two category members. As a result, a large majority of our participants' rules were more complex than those found with the train stimuli.

This study has shown that by activating an abstract feature space, theoretical knowledge provides certain biases that are used in rule induction. Because of these biases, the THEORY group appeared able to consider fewer rules that potentially describe a category than the STANDARD group. This might be the case for two reasons. First, some rules involving low-level features may be eliminated, since only a subset of low-level features will be relevant to an abstract hypothesis. Thus, a person in the THEORY group might not consider the features `hands at side` and `smiling` in a rule, since they would not be relevant to the abstract feature `detail`. Second, abstract hypotheses let people consider fewer competing rules during learning, since they can treat different features as equivalent. For example, a person in the THEORY group might view the features `has shoelaces` and `has buttons on jacket` as support for the single rule `drawings will be detailed`. In contrast, after examining a subset of category members that contain both features, someone in the STANDARD group might consider both features as distinct, potential rules for the category.

4. Study 2: Sequential Rule Induction

Although the first study demonstrated how theoretical knowledge guides rule induction, its methodology was limited in several important ways. Notably, people seldom acquire rules or concepts by examining a number of preclassified examples at the same time, as they did in the first study. In more realistic learning situations, rule induction is typically incremental and items are not preclassified. Rules are acquired gradually and modified in response to new examples. Second, although the previous methodology lets one assess whether or not theoretical and empirical knowledge interact, it does not clearly assess *how* they interact.

Our second study addresses this interaction by examining the development and modification of rules as training items are presented sequentially. We begin by giving people meaningful labels for the categories, then ask them to give us a "first-impression" rule of the category before they have seen any items. We assume that this rule primarily is based on a person's theories. We can then examine how the empirical evidence (i.e., training items and feedback), in conjunction with people's theories, causes people to modify their rules and affects the interpretation of succeeding training items. In this way, we can gain some insight into how empirical knowledge interacts with theoretical knowledge and vice versa. In this study we also examine in more detail the kinds of hypotheses that people formulate. We will show how different hypotheses (activated by different theories) dramatically affect the interpretation and selection of the particular features that people use in their rules.

4.1 Method

The subjects were 40 undergraduate students (male and female), attending the University of Illinois, who received course credit for participating in the study. A participant was randomly assigned to one of two groups of 20 subjects. Both groups learned about the same two categories (those shown in Figure 3). However, the groups were given different labels for the categories. One group (the FARM/CITY group) was told that the categories were drawings done by farm kids and drawings done by city kids. The other group (the CREATIVE/NONCREATIVE group) was told that the categories were drawings done by creative kids and drawings done by noncreative kids.

Students learned about the categories in the following manner. First, they were given the names of the categories. Then, before seeing any of the training items, they wrote down their initial rules for classifying the drawings. Next, students were shown the drawings one at a time, presented in a random order. A given drawing was not preclassified, and students had to decide to which category it belonged. After this decision, students wrote down the rule that they used to determine category membership. Then, they were given feedback on the correctness of their decision. After this feedback, students were given the option of modifying their rule. This process was repeated for each of the ten drawings. Afterward, students wrote down their overall rule for classifying drawings into the two categories.

4.2 Results

We discuss the results of this study in terms of two broad questions. First, how did theoretical knowledge affect the use and interpretation of the training data? Second, how did the training instances, in turn, affect the use of theoretical knowledge? Examining these questions provides some insight into how empirical and theoretical knowledge interact. Later, we sketch a model of this interaction and compare it to existing machine learning systems that have combined empirical and theory-based learning.

4.2.1 EFFECTS OF THEORETICAL KNOWLEDGE ON TRAINING DATA

Our results indicate that the two groups, given different labels, used different theories in learning about the same sets of drawings. In turn, these different theories had three important effects on how people processed the data. First, different theories caused people to selectively attend to different features in the training items. Second, and more striking, different theories caused people to interpret the same data differently. For example, people with different theories sometimes interpreted the same part of a drawing as a different feature. Third, like the first study, theories allowed people to view different features as similar at a higher level of abstraction.

Different labels activate different theories. To assess whether or not different labels activated different theories, we compared the initial rules of the two groups before any training items were presented. Differences in theoretical knowledge should be most apparent in these rules, since they were formulated in the absence of empirical evidence. In general, the initial rules of the groups were quite different, suggesting that they operated with different theories during the task. An examination of the protocols revealed that the initial rules of the CREATIVE/NONCREATIVE group were of two major types. Seventy percent of this group stated that creative kids would draw pictures with more detail and add more things to their drawings, whereas noncreative kids would draw pictures that were more basic and simple. For example, one subject stated that "creative kids will draw more detail — like eyelashes, teeth, curly hair, shading and coloring in. Noncreative kids draw more stick-figurish people." Twenty-five percent of this group stated that creative kids would draw pictures that were more imaginative and unusual than the pic-

tures drawn by noncreative kids. For example, one subject stated that "creative kids would stray from basics, show imagination — noncreative kids will draw regular, plain clothes, straight hair, basic facial features."

On the other hand, most of the FARM/CITY group stated that drawings would reflect the kinds of people that farm and city kids encountered in their environments and that these people could be identified by their clothing. Seventy-five percent of this group mentioned differences in the kind of clothing that the two types of drawing would have. For example, one subject stated that "farm kids will draw people with overalls, straw or farm hats. City kids will draw people with ties, suits."

Selective attention to features. It was clear from the protocols that the two groups selectively attended to different features in the input. Furthermore, we can relate these differences to the different theories of the groups. To show this, we examined two large classes of features that appeared in the rules. These classes were body terms like `arms`, `body`, `dancing`, `eyes`, `face`, `figure`, `hair`, `proportion`, `teeth`, and `running`, and clothing terms such as `belt`, `button`, `clothing`, `dress`, `jacket`, `pants`, `shoes`, `sleeves`, `suit`, `tie`, and `wearing`. For the two groups, we counted the number of different features (i.e., types) that people used in their rules as well as the number of occurrences (i.e., tokens) of each feature.

Table 2 shows the number of occurrences and different types of body terms and clothing terms listed by the two groups. The rules of the CREATIVE/NONCREATIVE group had almost three times as many occurrences of body terms (487 versus 173) and about twice as many types (60 versus 29) as the rules of the FARM/CITY group. In contrast, the rules of the FARM/CITY group had almost three times as many occurrences of clothing terms (295 versus 116) and slightly more types (25 versus 19) than the rules of the CREATIVE/NONCREATIVE group.

Why did the different theories cause the CREATIVE/NONCREATIVE group to consider mostly body features and the FARM/CITY group to consider mostly clothing features? One apparent reason for this difference is related to the different hypotheses that were initially activated by people's simple theories.[4] Recall that the FARM/CITY group hypothesized that farm and city children would draw people that they typically encountered in their different environments. Differences in the type of

4. The difference also is related to other hypotheses that the two groups formed over the course of category learning and not just to the initial hypotheses.

Table 2. Number of occurrences of body terms and clothing terms in Study 2.

	FARM/CITY	CREATIVE/NONCREATIVE
	Body Terms	
DIFFERENT TYPES	29	60
OCCURRENCES	173	487
	Clothing Terms	
DIFFERENT TYPES	25	19
OCCURRENCES	295	116

clothing (as opposed to body parts) would indicate whether the person was from a farm or a city. As a result, the FARM/CITY group attended more to clothing features than to body features. On the other hand, the use of body features provides plausible evidence for the two initial hypotheses entertained by the CREATIVE/NONCREATIVE group. Smaller-sized body parts (e.g., `eyebrows`, `teeth`, `fingers`) indicate that a drawing is detailed as opposed to simple (at least for a child's drawing). Similarly, references to body movement (e.g., `dancing`, `climbing`) provide evidence that a child's drawing is unusual as opposed to ordinary.

Interestingly, when the CREATIVE/NONCREATIVE rules included clothing items, they tended to refer to relatively small items of clothing (e.g., `shoelaces`, `buttons`, `belts`, `pockets`). In fact, the average size of clothing items mentioned by the CREATIVE/NONCREATIVE group was considerably less than that of the FARM/CITY group. Small clothing items provide evidence for detail in the drawings, one of the two initial hypotheses of the CREATIVE/NONCREATIVE group.

Interpreting the same data differently. We also examined the features that people mentioned about a given drawing. The most striking finding was that subjects sometimes interpreted the same part of a drawing as a different feature. Furthermore, it appeared that people's theoretical expectations determined the nature of these features. We can illustrate this process using examples from Figure 3. For drawing 5, a person in the CREATIVE/NONCREATIVE group referred to the part indicated by the arrow as `buttons`. The person mentioned this feature as evidence

of detail, which implied that the drawing was done by a creative child. On the other hand, a person in the FARM/CITY group interpreted the same part of the drawing as a tie. The person mentioned this feature as evidence that the drawing depicted a business-person, which implied that the drawing was done by a child from the city. As a second example, a person in the CREATIVE/NONCREATIVE group stated that drawing 3 of Figure 3 depicted someone dancing. This feature, in turn, showed imagination and implied that the drawing was done by a creative child. In contrast, a person in the FARM/CITY group interpreted the drawing as someone climbing in a playground. This feature, in turn, implied that the person in the drawing was from the city and therefore that the drawing was done by a child from the city. As a final example, a person in the CREATIVE/NONCREATIVE group interpreted the part of drawing 4 indicated by an arrow as a pocket. The person mentioned this feature as evidence of detail, which implied that the drawing was done by a creative child. In contrast, a person in the FARM/CITY group interpreted the same part of the drawing as a purse. This feature implied that the drawing depicted a city person and therefore was drawn by a child from the city.

Besides interpreting the same part of a drawing as a different feature, participants often interpreted a drawing as depicting different, stereotypical people. For example, in Figure 3, five people in the CREATIVE/NONCREATIVE group interpreted drawing 7 as a person from a culture different from that of America, such as a Frenchman, which implied that it was done by a creative child. In contrast, six people in the FARM/CITY group interpreted this drawing as a person from a city, such as a bellboy, which implied that it was done by a city child. No one in the CREATIVE/NONCREATIVE group mentioned that the drawing was done by a city person, and only one FARM/CITY subject mentioned that the drawing was done by a person from a different culture. As another example, in Figure 3, two subjects in the CREATIVE/NONCREATIVE group mentioned that drawing 9 was a regular, ordinary person and therefore was done by a noncreative child. In contrast, three subjects in the FARM/CITY group mentioned that the drawing was a farmer or cowboy and therefore was done by a farm child.

This type of theory-based processing was much more common in the FARM/CITY group (58 examples) than in the CREATIVE/NONCREATIVE group (21 examples). Recall that the major initial expectation of the FARM/CITY group was that children from the farm or city would draw

people from those environments. Not surprisingly, this group interpreted a number of drawings as depicting people typically found in such settings. People from the city included a `cool guy from a gang`, a `bellboy`, a `child climbing on a playground`, a `bus driver`, and a `professional`. People from the farm included a `farmer`, a `cowboy`, a `carpenter`, and a `blue-collar craftsman`.

4.2.2 THE EFFECT OF DATA ON THEORETICAL KNOWLEDGE

The next issue that we address is how training data caused people to revise their theories about the drawings of creative and noncreative children. Specifically, we will examine how rules are affected by training items that have been incorrectly classified. Recall that after classifying a drawing and writing down their rule, participants were given feedback on their decisions. At this point, participants had the option of modifying their rule. In general, when people were given positive feedback, they did not modify their rules. However, when given negative feedback, subjects used a number of strategies to modify their rules. Below, we describe three of these strategies and provide examples of each.

Using theory-based features in context-specific ways. At times people altered their use of a theory-based feature (e.g., `detail`), depending on the context. In particular, they mentioned that a feature was true of a drawing and constituted evidence for one of the categories. However, given feedback that the drawing belonged to the contrasting category, they *changed their criteria* for applying that feature. For example, one person mentioned that a drawing depicted detailed clothing and therefore that it was drawn by a creative child. But when told that the drawing was done by a noncreative child, the person changed his or her rule to "drawings done by creative children would be *more* detailed." In this example, the person adjusted the theory-based belief that creative children will show detail. Specifically, for a drawing to be classified as creative, it must have an increased amount of `detail` relative to the noncreative drawing just presented.

As another example, one person stated that a drawing was done by a city child because "it looks very detailed, has colored-in places." However, upon being told that the drawing was done by a farm child, the person reexamined the details of the drawing and stated that "drawings with detail in specific clothing is more of a rule for city kids — not

detail in body movement as this one had." As in the first example, the person has altered his or her use of the theory-based feature `detail`. Specifically, for a drawing to be classified as done by a city child, it must now show detailed clothing as opposed to detail in body movement.

Reinterpreting features to preserve theory-based beliefs. Given feedback that they had incorrectly classified a drawing, people sometimes reinterpreted the features that they had used in making that incorrect decision. Then, they used these reinterpreted features as evidence for the correct category. This strategy might allow people to preserve the validity of their theory-based beliefs. For example, one person mentioned that a drawing was done by a city child because it depicted a television character, and city children watch more television. However, when told that the drawing was done by a farm child, the person reinterpreted the character as one created from a farm child's imagination. Such a strategy would let the person's theory-based belief (namely, that city children watch more television) remain intact.

As another example, one person thought that the clothing in a drawing was a `city uniform` and was drawn by a city child. Given feedback that the drawing was done by a farm child, the person reinterpreted the clothing as a `farm uniform`. As before, such a strategy allows the person's theory-based beliefs to remain intact. That is, after modifying the rule, the person's belief that city children would draw people with city uniforms remains a valid hypothesis.

Shifting to evidence that supports alternative theory-based beliefs. Given feedback that they had incorrectly classified a drawing, people often considered alternative evidence for the correct category. For example, one person mentioned "stiff arms, misproportioned legs, and feet stemming out in both directions" as evidence that a drawing was noncreative. Told that the drawing was creative, the person stated that "avoiding those attributes, there is some eye for detail in the face." As a second example, one person classified a drawing as done by a city child because it depicted a "child climbing in a playground." Told that the drawing was done by a farm child, the person noted that the drawing also showed "plaid clothing." In fact, some participants in the FARM/CITY group believed that plaid clothing was typical of people from rural areas. As a third example, one person categorized a drawing as noncreative because it was not "really real in the way the head and arms don't look like they

belong to the legs." Given negative feedback, the person suggested that creative drawings have "a lot of effort put into them" and that this drawing included "facial features and extra things (belt, purse)."

4.3 Discussion

These results suggest some of the ways that theoretical expectations and empirical evidence interact. They also suggest that the problem of induction should be viewed from a somewhat different perspective than that typically seen in disciplines of cognitive psychology and machine learning. In particular, our results imply that theoretical expectations and empirical evidence can interact closely to determine the features that induction operates upon. The major evidence for this claim comes from the finding that people with different theoretical expectations sometimes interpreted the same parts of drawings as different features, or the same drawing as representing a different kind of person. Thus, theoretical knowledge affected feature interpretation. On the other hand, people clearly did not just "see what they wanted to see." The data provided constraints that allowed some interpretations and not others. For example, consider the part indicated by the arrow in drawing 4 of Figure 3. Although one subject interpreted the configuration as a pocket, therefore providing evidence of detail, the person did not interpret the configuration as shoelaces, eyebrows, or buttons, even though such features also provided evidence for detail. In the following section, we outline a speculative model of how theory-based beliefs and perceptual input interact to determine features.

Besides interacting to constrain features, theoretical expectations and empirical evidence can interact to produce the explanations that participants gave as rules. Often it was clear that an explanation for one drawing's membership in a category was affected by previous items and feedback. There were a number of ways that such information influenced the current explanation. Although not described in the results, people often constructed explanations that were similar to ones used in explaining correct classifications. They also avoided explanations that had been used in explaining incorrect classifications. These findings simply suggest that people were sensitive to previous explanations which had been successful or unsuccessful. As described in the results, people also adjusted their explanations when given negative feedback. They sometimes readjusted the criteria for their applicability or reinterpreted the features upon which they based those explanations.

5. Models of Theoretical and Empirical Learning

We have suggested that people learn about categories by combining theory-based beliefs with empirical evidence. As mentioned in the introduction, researchers in machine learning have developed a number of systems that combine data-driven or empirical learning with theory-driven or explanation-based learning. These models can be divided into two broad classes. Below, we describe these classes and discuss their plausibility, given the current findings.

5.1 Previous Models

Many models learn by noting similarities among the training examples and passing the output to a theory-driven component. Some of these systems, like OCCAM (Pazzani, 1987), use an empirical learning component to acquire domain theory knowledge, which can then be used by the explanation-based learning component to construct explanations. Basically, these methods improve the domain theories used by explanation-based learning systems. For example, Lebowitz (1986b) describes a system that employs UNIMEM to detect similarities across a number of training examples and then uses explanation-based learning to explain them. This approach apparently improves the efficiency of the explanation process.

In contrast, some models carry out explanation-based learning followed by empirical learning. These systems first proceed by processing training examples in a theory-driven manner and then sending output to a data-driven component (e.g., Cohen, 1988; Flann & Dietterich, 1989; Kedar-Cabelli, 1985; Mooney & Ourston, 1989; Pazzani, 1988; Shavlik & Towell, 1989; Yoo & Fisher, this volume). For example, Mooney and Ourston's (1989) IOU (induction over the unexplained) first applies explanation-based learning to training items. Features of the items that do not enter into the explanations are input to an empirical learning component, which detects features that tend to occur across the items. As previously mentioned, this approach is important because many concepts have both explanatory and nonexplanatory components, whereas the theoretical knowledge of explanation-based systems is typically explanatory in nature. Other systems use explanation-based learning to construct explanation trees for a number of training items and employ empirical methods to find the largest subtree that is common to all

the explanations (e.g., Flann & Dietterich, 1989; Yoo & Fisher, this volume).

In both types of models, the interaction between empirical and explanation-based learning is indirect and typically operates in one direction. In a number of these systems, the first module acts as a *filter* for the input to the second module by reducing the number of features that the second processes. For example, the explanation-based module in IOU first processes the training items and then passes a subset of their features (i.e., the unexplained features) to the empirical learning module. In Lebowitz's system, the empirical component processes training items and passes the common features to the explanation-based module. In these examples, one component influences the other but not vice versa.

However, our results suggest that the interaction between theory-driven and empirical learning can be much more direct. In this case, the learning modules are tightly coupled, with one module's processing "interwoven" with that of the other. Both modules influence each other directly. Now we turn to a model that incorporates this idea.

5.2 A Tightly Coupled Approach

We propose an alternative model, in which an explanation or rule is a function of three sources of knowledge: prior theories, explanations and data associated with previous training items, and the data provided by the current training item. In contrast, current models that integrate explanation-based and empirical learning formulate explanations that are constructed from *two* sources of knowledge: prior theories and the data provided by the current item. Below we speculate on the characteristics of this model and compare them to those current systems that combine explanation-based learning and empirical learning.

5.2.1 CONSTRUCTING DOMAIN RULES

In learning about categories of children's drawings, it is unlikely that people have a domain theory that can be applied directly to such drawings. For example, although many people might believe that creative drawings will show detail, what counts as evidence for detail might vary widely with the drawer. For a creative child, **shoelaces**, **buttons**, and **eyebrows** might provide good evidence. However, these features probably are not good indicators of detail in the drawings of creative adults.

Instead, people probably use general knowledge and functional specifications of categories to *construct hypothetical domain rules*, which they attempt to verify in the course of learning about the category. In the case of children's drawings, people probably access general knowledge structures corresponding to creativity, children, drawing, and what a typical person looks like. These structures are called *schemata* or *frames* (e.g., Minsky, 1975; Palmer, 1978; Norman & Rumelhart, 1975). Schemata capture (among other things) people's stereotypical beliefs and prior probabilities about events and objects, the visual appearance of objects, the temporal sequence of actions within an event, and so on.

Given the functional specification of a category (e.g., the meaningful label drawings done by creative children) and these basic knowledge structures, a person may plausibly construct an initial hypothesis about the category. In contrast to this rule-construction process, an explanation-based learning system typically has a domain theory containing preexisting rules. Using these rules, it directly links the features of a training item to the functional specification of a category. An exception to this requirement is Pazzani's (1987) OCCAM, which learns new domain theory rules from very high-level knowledge (e.g., about causality) much like our studies indicate that humans learn.

5.2.2 FEATURE INTERPRETATION

Our results also suggest that part of learning involves feature interpretation. This process can occur through a complex interaction of top-down knowledge (i.e., the hypothetical rules) and bottom-up knowledge (i.e., data from training items). There are several ways in which these types of knowledge might mutually influence each other during feature interpretation. First, theoretical expectations may initially activate feature schemata, and data from the current training item may constrain which schemata are plausible. These data include low-level shape descriptors and the relations among them, which are delivered by the perceptual system (Biederman, 1985; Marr & Nishihara, 1978).

Feature interpretation is a matter of finding a good match between feature schemata and the low-level data. As a simple example, consider again the part indicated by the arrow in drawing 4 of Figure 3. The hypothesis "creative children will show detail" will activate a wide range of schemata (e.g., button, shoelace, pocket, eyebrows, teeth). How-

perceived many of the same (concrete) features in the drawings. Clearly, low-level perceptual information can activate features in the absence of prior expectations. Furthermore, one type of constraint may dominate the other. For example, no matter how strong your top-down expectations were for seeing elephants, you would still perceive drawing 1 of Figure 3 as a person.

5.2.3 CONTEXT-SENSITIVE EXPLANATIONS

In our model, people's explanations are sensitive to empirical evidence involved in the learning task. In particular, feedback and hypothesized features associated with previous examples influence the construction of later explanations. Whereas people construct context-sensitive explanations, systems that have used explanation-based learning followed by empirical learning have typically constructed context-independent explanations. That is, an explanation of an example E_i does not affect an explanation of the next example E_{i+1}.

We can think of several reasons why context sensitivity of explanations is a desirable quality, especially in the case where a learner's theoretical expectations or domain theory is weak. First, if an explanation of an example is incorrect, it would be sensible to avoid constructing a similar explanation for another item (i.e., "don't make the same mistake twice"). The incorrect explanation may indicate that the learner's theoretical expectations or domain theory needs modification. Likewise, if the same explanation is constructed for a number of examples, it might be sensible to prefer constructing that explanation for a new item. The successful explanation may indicate that aspects of the learner's theoretical expectations or domain theory are especially accurate.

Another important reason to prefer context-sensitive explanations is that some aspects of knowledge are *relative*. When applied to an item, the meanings of features such as `detailed`, `tall`, `heavy`, and `small` are defined with respect to other objects (e.g., Rips & Turnbull, 1980). For example, a small horse is small relative to other horses, but it is not small relative to many other objects (e.g., frogs and pencils). The fact that knowledge can be relative means that some explanations crucially interact with training data. For example, as noted in the second study, an explanation involving relative features may essentially be correct, but need to be "adjusted" based on the empirical evidence. Thus, one person believed that the drawings done by creative children would be

detailed, but after examining a drawing done by a noncreative child, the person adjusted the criteria for calling a drawing detailed. Here, the explanation remained basically the same, but was adjusted based on the empirical evidence.

One dramatic example of this interaction occurs with the use of the feature detailed. Consider three (hypothetical) sets of faces, ordered by increasing amount of detail. The first set (and least detailed) was drawn by noncreative children. The second set (and the second most detailed) was done by creative children. The third (and most detailed) was drawn by creative adults. The explanation "drawings will show *more* detail" accounts for why drawings are done by creative children when compared to noncreative children, but the same explanation will not account for why drawings are done by creative children when compared to creative adults. In fact, the opposite explanation that "drawings will show *less* detail" might be more appropriate.

6. Summary

In this chapter we presented two studies that examined the roles of theoretical expectations and empirical evidence in rule induction. These studies produced several major findings.

First, theoretical expectations strongly affect the kinds of rules that people construct for a given category. These rules are qualitatively different from those constructed by people without such expectations. Theoretical expectations appear to activate hypotheses about abstract features that might be true of a category. People who had these expectations tended to form rules with abstract features and to operationalize these abstract features by linking them to more concrete features in the training examples. In contrast, people without such expectations seldom used abstract features or operationalization in their rules. Instead, they focused on finding a set of simple, observable features that described a category. It also appears that theoretical expectations promote generalization across features. People operating with such expectations were able to treat different features equivalently. Therefore, they were able to find commonalities among category members that people without such expectations were unable to find.

Second, theoretical knowledge closely interacts with empirical evidence, and these types of knowledge mutually influence each other. In particular, the two sources of knowledge interact to determine hypo-

thetical features upon which induction operates. They also mutually influence the types of explanations that people construct for a category. We briefly described how a learning system might closely integrate these sources of knowledge.

The view that people's theoretical expectations closely interact with experience generally has not been emphasized in machine learning models that integrate these two sources of knowledge. Furthermore, the notion that theoretical and empirical knowledge jointly determine the features for the induction process has been largely ignored in both cognitive psychology and machine learning. We hope that this paper contributes to a clearer understanding of these very complex problems.

Acknowledgements

We gratefully acknowledge the comments of Woo-kyoung Ahn, Douglas Fisher, Evan Heit, Pat Langley, Colleen Seifert, Edward Smith, and Rick Riolo on a previous version of this chapter. We also thank Bob Dylan for inspiration.

References

Adelman, L. (1981). The influence of formal, substantive, and contextual task properties on the relative effectiveness of different forms of feedback in multiple-cue probability learning tasks. *Organizational Behavior and Human Performance*, *27*, 423–442.

Ahn, W., & Brewer, W. F. (1988). Similarity-based and explanation-based learning of explanatory and nonexplanatory information. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 50–57). Montreal, Quebec: Lawrence Erlbaum.

Biederman, I. (1985). Human image understanding: Recent research and a theory. *Computer Vision, Graphics, and Image Processing*, *32*, 29–73.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: John Wiley & Sons.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Collins, A. (1978). *Human plausible reasoning* (Tech. Rep. No. 3810). Cambridge, MA: Bolt, Beranek, & Newmann.

Collins, A., & Michalski, R. S. (1989). The logic of plausible reasoning: A core theory. *Cognitive Science, 13*, 1–50.

DeJong, G. (1988). An introduction to explanation-based learning. In H. E. Shrobe (Ed.), *Exploring artificial intelligence*. San Mateo, CA: Morgan Kaufmann.

DeJong, G., & Mooney, R. (1986). Explanation-based learning: An alternative view. *Machine Learning, 1*, 145–176.

Dietterich, T. G., London, B., Clarkson, K., & Dromey, G. (1982). Learning and inductive inference. In P. R. Cohen & E. A. Feigenbaum (Eds.), *The handbook of artificial intelligence*. San Mateo, CA: Morgan Kaufmann.

Ellman, T. (1989). Explanation-based learning: A survey of programs and perspectives. *Computing Surveys, 21*, 163–221.

Flann, N. S., & Dietterich, T. G. (1989). A study of explanation-based methods for inductive learning. *Machine Learning, 4*, 187–226.

Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning, 2*, 139–172.

Fisher, D. H. (1988). A computational account of basic level and typicality effects. *Proceedings of the Seventh National Conference on Artificial Intelligence* (pp. 233–238). Saint Paul, MN: Morgan Kaufmann.

Fisher, D. H., & Langley, P. (1985). Approaches to conceptual clustering. *Proceedings of the Ninth International Conference on Artificial Intelligence* (pp. 691–697). Los Angeles, CA: Morgan Kaufmann.

Goodenough, F. L., & Harris, D. B. (1950). Studies in the psychology of children's drawings II, 1928–1949. *Psychological Bulletin, 47*, 363–433.

Hanson, S. J., & Bauer, M. (1989). Conceptual clustering, categorization, and polymorphy. *Machine Learning, 3*, 343–372.

Harris, D. B. (1963). *Children's drawings as measures of intellectual maturity*. New York: Harcourt Brace & World.

Haygood, R. C., & Bourne, L. E. (1965). Attribute and rule-learning aspects of conceptual behavior. *Psychological Review, 72*, 175–195.

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

Kedar-Cabelli, S. (1985). Purpose-directed analogy. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society* (pp. 150–159. Irvine, CA: Lawrence Erlbaum.

Keil, F. C. (1981). Constraints on knowledge and cognitive development. *Psychological Review*, *88*, 197–227.

Koppitz, E. M. (1984). *Psychological evaluation of human figure drawings by middle school pupils*. Orlando, FL: Grune and Stratton.

Langley, P. (1987). Machine learning and concept formation. *Machine Learning*, *2*, 99–102.

Lebowitz, M. (1986a). Not the path to perdition: The utility of similarity-based learning. *Proceedings of the Fifth National Conference on Artificial Intelligence* (pp. 533–538). Philadelphia, PA: Morgan Kaufmann.

Lebowitz, M. (1986b). Integrated learning: Controlling explanation. *Cognitive Science*, *10*, 219–240.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London B 200* (pp. 269–294).

Medin, D. L. (1983). Structural principles of categorization. In B. Shepp & T. Tighe (Eds.), *Interaction: Perception, development, and cognition*. Hillsdale, NJ: Lawrence Erlbaum.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. Cambridge: Cambridge University Press.

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

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

Medin, D. L., Wattenmaker, W. D., & Michalski, R. S. (1987). Constraints and preferences in inductive learning: An experimental study of human and machine performance. *Cognitive Science*, *11*, 299–339.

Michalski, R. S. (1983a). A theory and methodology of inductive learning. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (Vol. 1). San Mateo, CA: Morgan Kaufmann.

Michalski, R. S. (1983b). A theory and methodology of inductive learning. *Artificial Intelligence, 20*, 111–161.

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The psychology of computer vision*. New York: McGraw-Hill.

Mitchell, T. M. (1980). *The need for biases in learning generalizations* (Tech. Rep. No. CBM-TR-117). New Brunswick, NJ: Rutgers University, Department of Computer Science.

Mitchell, T. M. (1982). Generalization as search. *Artificial Intelligence, 18*, 203–226.

Mitchell, T. M., Keller, R. M., & Kedar-Cabelli, S. T. (1986). Explanation-based generalization: A unifying view. *Machine Learning, 1*, 47–80.

Mooney, R. J., & Ourston, D. (1989). Induction over the unexplained: Integrated learning of concepts with both explainable and conventional aspects. *Proceedings of the Sixth International Workshop on Machine Learning* (pp. 5–7). Ithaca, NY: Morgan Kaufmann.

Mostow, J. (1983). Machine transformation of advice into a heuristic search procedure. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (Vol. 1). San Mateo, CA: Morgan Kaufmann.

Muchinsky, P. M., & Dudycha, A. L. (1974). The influence of a suppressor variable and labeled stimuli on multiple cue probability learning. *Organizational Behavior and Human Performance, 12*, 429–444.

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

Norman, D. A., & Rumelhart, D. E. (1975). Memory and knowledge. In D. A. Norman & D. E. Rumelhart (Eds.), *Explorations in cognition*. San Francisco: Freeman.

Nosofsky, R. M. (1986). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 87–108.

Palmer, S. E. (1975). Visual perception and world knowledge: Notes on a model of sensory-cognitive interaction. In D. A. Norman & D. E. Rumelhart (Eds.), *Explorations in cognition*. San Francisco: Freeman.

Pazzani, M. J. (1987). Explanation and generalization based memory. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society* (pp. 323–328). Irvine, CA: Lawrence Erlbaum.

Pazzani, M. J. (1988). Integrated learning with incorrect and incomplete theories. *Proceedings of the Fifth International Conference on Machine Learning* (pp. 291–297). Ann Arbor, MI: Morgan Kaufmann.

Pazzani, M. J., & Schulenburg, D. (1989). The influence of prior theories on the ease of concept acquisition. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 812–819). Ann Arbor, MI: Lawrence Erlbaum.

Pazzani, M. J., Dyer, M., & Flowers, M. (1987). Using prior theories to facilitate the learning of new causal theories. *Proceedings of the Tenth International Joint Conference on Artificial Intelligence* (pp. 277–279). Milano, Italy: Morgan Kaufmann.

Quinlan, J. R. (1983). Learning efficient classification procedures and their application to chess end games. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (Vol. 1). San Mateo, CA: Morgan Kaufmann.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81–106.

Rajamoney, S. A. (1986). *Automated design of experiments for refining theories* (Tech. Rep. No. UILU-ENG-86-2213). Urbana: University of Illinois, Department of Computer Science.

Rendell, L. (1985). Substantial constructive induction using layered information compression: Tractable feature formation in search. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence* (pp. 650–658). Los Angeles: Morgan Kaufmann.

Rips, L. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. Cambridge: Cambridge University Press.

Rips, L. J., & Turnbull, W. (1980). How big is big? Relative and absolute properties in memory. *Cognition*, *8*, 145–174.

Scott, P. D. (1987). *On the systematics of machine learning*. Unpublished manuscript, Department of Computer Science, University of Michigan, Ann Arbor.

Schank, R. C., Collins, G. C., & Hunter, L. E. (1986). Transcending inductive category formation in learning. *Behavioral and Brain Sciences*, *9*, 639–686.

Shavlik, J. W., & Towell, G. G. (1989). *Combining explanation-based and neural learning: An algorithm and empirical results* (Tech. Rep. No. 859). Madison: University of Wisconsin, Department of Computer Sciences.

Smith, E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Utgoff, P. E. (1986). A shift in bias for inductive concept learning. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (Vol. 2). San Mateo, CA: Morgan Kaufmann.

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

Wattenmaker, W. D., Nakamura, G. V., & Medin, D. L. (1987). Categorization processes and causal explanation. In D. Hilton (Ed.), *Contemporary science and natural explanation: Commonsense concepts of causality*. Sussex, England: Harvester Press.

Winston, P. H. (1975). Learning structural descriptions from examples. In P. H. Winston (Ed.), *The psychology of computer vision*. New York: McGraw-Hill.

Wisniewski, E. J, & Medin, D. L. (1991). *Is it a mornek or a plapel? Prior expectations about functionally-relevant features in category learning*. Unpublished manuscript, Department of Psychology, University of Michigan, Ann Arbor.

Wisniewski, E. J., Winston, H. A., Smith, R. G., & Kleyn, M. (1987). A conceptual clustering program for rule generation. *International Journal of Man-Machine Studies*, *27*, 295–313.

Wright, J. C., & Murphy, G. L. (1984). The utility of theories in intuitive statistics: The robustness of theory-based judgments. *Journal of Experimental Psychology: General*, *113*, 301–322.