On the Interaction of Theory and Data in Concept Learning

EDWARD J. WISNIEWSKI AND DOUGLAS L. MEDIN

Northwestern University

Standard models of concept learning generally focus on deriving statistical properties of a category based on data (i.e., category members and the features that describe them) but fail to give appropriate weight to the contact between people's intuitive theories and these data. Two experiments explored the role of people's prior knowledge or intuitive theories on category learning by manipulating the labels associated with the category. Learning differed dramatically when categories of children's drawings were meaningfully labeled (e.g., "done by creative children") compared to when they were labeled in a neutral manner. When categories are meaningfully labeled, people bring intuitive theories to the learning context. Learning then involves a process in which people search for evidence in the data that supports abstract features or hypotheses that have been activated by the intuitive theories. In contrast, when categories are labeled in a neutral manner, people search for simple features that distinguish one category from another. Importantly, the final study suggests that learning involves an interaction of people's intuitive theories with data, in which theories and data mutually influence each other. The results strongly suggest that straightforward, relatively modular ways of incorporating prior knowledge into models of category learning are inadequate. More telling, the results suggest that standard models may have fundamental limitations. We outline a speculative model of learning in which the interaction of theory and data is tightly coupled. The article concludes by comparing the results to recent artificial intelligence systems that use prior knowledge during learning.

The research was supported in part by the National Institute of Child Health and Development under fellowship award 5 F32 HD07279-02 given to Edward J. Wisniewski and by the National Science Foundation under grant 9110245 given to Douglas L. Medin. We thank Miriam Basock, Lee Brooks, Pat Cheng, Jeff Ellman, Deidre Gentner, Evan Heit, Arthur Markman, Lance Rips, Colleen Seifert, Edward Smith, and especially Pat Langley for trenchant comments on previous versions of this article. We also thank Leigh Elkins for assistance in running the experiments. Informal descriptions of some of these experiments appeared in Wisniewski and Medin (1991, 1994).

Correspondence and requests for reprints should be sent to Edward J. Wisniewski, Northwestern University, Department of Psychology, Swift Hall, 2029 Sheridan Rd., Evanston, IL 60208-2710.
INTRODUCTION

It seems obvious that our understanding of many categories is importantly determined by experiences with examples of these categories. In fact, a popular view in both cognitive psychology and artificial intelligence (AI) is that concept representations are simple functions of the properties of category instances. For example, prototype abstraction theories use examples to derive the central tendency of the category. As another example, discrimination net models use properties of examples to conduct a series of tests that act as a classification rule. This general type of learning has been given a variety of names, including: similarity-based learning (e.g., Lebowitz, 1986), empirical learning, and data-driven learning (e.g., Langley, 1981). According to this view, concept learning involves computations over examples.

Although experience is certainly important, empirical approaches to learning fail to give appropriate weight to the contact between previous knowledge and data in learning. Intuitively, it appears that people often draw upon their prior knowledge when learning about categories in the world. For example, consider the first time you encountered a computer. It is unlikely that you learned about computers without relying on any prior knowledge. You had probably read or been told that computers were used for text writing and programming. Also, seeing a computer for the first time may have activated your knowledge about televisions, typewriters, and calculators. Undoubtedly, this knowledge influenced your learning about computers. At the very least it probably determined which features you thought were relevant to computing. In general, it seems that learning must involve contact with people's prior knowledge. Several areas of research illustrate this point.

First, many researchers have suggested that people's concepts are richly interconnected and embedded in people's intuitive theories of the world (Carey, 1985; Keil, 1989; Medin & Wattenmaker, 1987; Miller & Johnson-Laird, 1976; G. Murphy, 1993; G. Murphy & Medin, 1985; Nisbett & Ross, 1980; Schank, Collins, & Hunter, 1986; C. Smith, Carey, & Wiser, 1985). For example, the concepts “money,” “bank,” and “pay” might be interconnected and embedded in theories of social and economic interaction. This view necessarily implies that when a person acquires a new concept, it must somehow be integrated with prior knowledge. So, learning the concept “loan” may require linking “loan” to the economic concepts and theories just mentioned.

Second, a number of researchers have shown that the degree to which prior knowledge is consistent with the structure of the category affects the learning of that category. Wattenmaker, Dewey, T. Murphy, and Medin (1986) examined how prior expectations affected the learning of categories that were either linearly or non-linearly separable (see, also, Nakamura, 1985). Two categories are linearly separable if they can be partitioned on the
basis of an additive function of weighted features that describe the categories. They found that which type of abstract category structure was easier to learn depended critically on the kinds of expectations induced in subjects. In particular, when expectations suggested that each property of an example was relevant to its function, then linearly separable categories were easier to learn. When expectations suggested that conjunctions of features described the category members, then nonlinearly separable categories were easier to learn.

As another example, Pazzani (1991) showed that prior expectations can facilitate the learning of disjunctive rules and interfere with the learning of conjunctive rules. Without prior expectations, people learn conjunctive rules more easily than disjunctive rules (Bruner, Goodnow, & Austin, 1956; Medin, Wattenmaker, & Michalski, 1987). In one experiment, Pazzani (1991) showed subjects pictures of people interacting with balloons and asked them to find a rule that predicted whether or not the balloons would be inflated. In one condition, the disjunctive rule “age of person is adult or action of person is stretching” characterized those pictures of balloons that were likely to be inflated. In another condition, the conjunctive rule “size of balloon is small and color of balloon is yellow” characterized these pictures. Subjects learned the disjunctive rule more easily than the conjunctive rule, as it was consistent with their prior expectations about the likelihood of a balloon being inflated. Furthermore, subjects learned the disjunctive rule more easily than a control group that did not have these expectations. In contrast, subjects learned the conjunctive rule less easily than a corresponding control group.

Finally, G. Murphy and Wisniewski (1989) showed that categories that violated people’s prior expectations about feature correlations were more difficult to learn than those that were consistent with such expectations. So, a novel type of animal that “eats wheat” and “lives in the water” was harder to learn about than one that “eats fish” and “lives in the water.”

It seems clear that concept learning is a process involving both experiences and prior knowledge. In general, however, most psychological models of concept learning have ignored the role of prior knowledge or expectations and have focused solely on the role of experience (e.g., J.R. Anderson, 1990; Busemeyer & Myung, 1988; Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Gluck & Bower, 1988; Heit, 1992; Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1988, 1991). The situation is somewhat different in AI where a number of researchers have developed explanation-based learning systems, which use prior knowledge to acquire concepts (e.g., Dejong & Mooney, 1986; Flann & Dietterich, 1989; Mitchell, Keller, & Kedar-Cabelli, 1986; Pazzani, 1985). Nonetheless, many AI learning systems focus primarily on the role of experience (e.g., Dietterich, London, Clarkson, & Dromey, 1982; D. Fisher, 1987; Michalski, 1983a, 1983b; Mitchell, 1982; Pagallo, 1989; Quinlan, 1983, 1986; Sutton & Matheus, 1989).
This article examines how prior knowledge influences learning. In particular, it explores three possible models of prior knowledge effects, which are illustrated in Figure 1. The top of the figure shows a sequence of three training items (belonging to one of two categories). The rest of the figure shows
the output of each system after applying its prior knowledge to a training item and then applying induction. The goal of a model is to learn a classification rule for Category 1. According to the *selection* model, knowledge first weights or *selects* features and then a version of the *standard model* of concept learning (described later) focuses on these selected features. Pazzani's (1991) POSTHOC and Lien and Cheng's (1989) framework for causal induction are examples that fit this general approach. G. Murphy (1993) has also suggested that knowledge functions primarily to select the relevant features for learning. According to the *addition* model, knowledge not only selects features but also allows one to infer additional, relevant features. A version of the standard model of concept learning then focuses on the selected and added features. In machine learning, there are a number of systems that are variations of the addition model (e.g., Flann & Dietterich, 1989; Mooney & Ourston, 1989; Yoo & D. Fisher, 1991). These systems typically combine a knowledge-based addition and selection component with an empirical component.

It is important to notice that the selection and addition approaches are relatively independent of the learning process. That is, knowledge first selects and/or adds features to category items. Learning then operates over the results of this process. In this sense, knowledge effects and learning involve *independent* modules, with one process sending its results to the other.

We shall argue that this approach is inadequate and that the interaction of knowledge and learning is more *tightly coupled* than this view allows. The third model shown in Figure 1 illustrates an example of this tightly coupled approach. Initially, prior knowledge was applied to the first training item to infer a feature $F_1$ from features $F_1$ and $F_2$, and a feature $F_5$ from features $F_4$ and $F_6$. Learning then focuses on these features. Later in the sequence, however, when the system applies prior knowledge to the third training item, it infers the feature $F_5$ from the features $F_4$ and $F_6$ (rather than from $F_4$ and $F_5$ as it did in the second training item). Here the model's prior knowledge ($F_4$ and $F_6 = > F_5$) is modified by experience and this modified knowledge ($F_5$ and $F_6 = > F_5$) will then exert its effects on future learning. This example illustrates the critical point that learning may not only determine how features are associated with a category but may also involve learning of the *features themselves*. In this example, the model initially hypothesized that feature $F_5$ characterized the second item in the sequence. However, after experiencing the sequence, the system *redefines* its notion of feature $F_5$. This feature *no longer characterizes* the second item.

We begin by describing a broad framework with which to view many models of concept learning. Within this framework, we will suggest that at a certain level of abstraction, these models share important assumptions. These assumptions form the basis of what we call the "standard model of concept learning."

Next, we discuss our approach to studying the effects of
prior knowledge on learning. We then present experiments that evaluate the selection and addition models and suggest that neither model will fully capture knowledge effects. We summarize by questioning the general validity of standard models of concept learning (i.e., those that do not use prior knowledge) and by outlining a less traditional view of learning in which knowledge and experience closely interact and assess its generality. Finally, the role of knowledge in human learning is compared to the treatments of knowledge effects in AI learning programs.

The Standard Model of Concept Learning
One can view models of concept learning as consisting of a number of very basic components. A given model receives some type of input, in the form of training items. Using a learning algorithm, it constructs a concept or representation of a given category. Each model typically has a matching function that uses the concept to determine the category in which a novel item belongs.

Models of concept learning also share some general assumptions about these components. First, the models assume that their input involves a space of prespecified, unambiguous features. Typically, the training items consist of a fixed set of attribute-value pairs. Sometimes the items are simple visual stimuli that are easily mapped onto attribute values. A second assumption is that the learning algorithm is a syntactic process that involves selecting an appropriate subset of features from this space. Third, this selection process is primarily driven by statistical properties of features (it typically involves the relative frequency of features within a category and its contrast categories). This process may also weight these features by perceptual salience and/or by frequency. A consequence of this assumption is that features are only important to the extent that they are diagnostic of category membership. Fourth, classification is based on a matching function that uses syntactic identity to determine which features of a novel item match and mismatch those in the concept. Classifying an item as a member or non-member of a category is then some function of the feature matches and mismatches. Table 1 summarizes these assumptions.

The standard model may be illustrated by way of an example. E. Smith and Medin (1981) described a prototype model they suggested is a more general version of a number of existing models (e.g., Collins & Loftus, 1975; Hampton, 1979; McCloskey & Glucksberg, 1979). Input to the model
consists of learning items that are described as lists of attribute-value pairs. The learning algorithm involves selecting the modal value of each attribute from the among these items. (The algorithm also weights these features by some function of their salience and relative frequency among the category members.) The result of this selection process is a concept consisting of a subset of the features present in the training items. Deciding whether a novel item belongs to a category first involves determining those features of the item that are syntactically identical to those present in the concept. The matching function computes a weighted sum of these features and the item is considered a member of the category if the sum exceeds a threshold.

There are a number of classes of models that subscribe to the standard view, although they typically differ in their learning algorithm and matching function. In addition to prototype models, they include rule-based models (e.g., Bruner et al., 1956; Haygood & Bourne, 1965; Quinlan, 1983, 1986), exemplar models (e.g., Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1986a, 1986b, 1988), and recent models of incremental conceptual clustering (e.g., Ahn & Medin, 1992; J.R. Anderson, 1990; Cheeseman et al., 1988; D. Fisher, 1987; Hanson & Bauer, 1989). In general, these models have been quite successful in accounting for a variety of categorization phenomena, particularly in knowledge-poor domains where the feature structure is straightforward (e.g., J.R. Anderson, 1990). Indeed, it is the prior success of these models that motivates our attempt to extend them to knowledge-rich contexts. Although there are important differences among these models, which have consequences for their ability to model psychological findings, this article only focuses on the general assumptions that all of these models make.

If prior knowledge functions primarily to add or select features (as illustrated by the first two models in Figure 1) then incorporating prior knowledge into any of these models will be straightforward. If one can specify the prior knowledge that selects features, or the knowledge that allows features to be inferred from other features, then this knowledge would just be another input to the learning algorithm. On the other hand, if the interaction of knowledge and experience is tightly coupled during learning (as illustrated by the third model in Figure 1), then the learning algorithm itself will have to be modified. To the extent that prior knowledge enters into learning and is tightly coupled with experience, it is important to study learning and prior knowledge effects together.

**Studying Knowledge Effects in Category Learning**

The typical paradigm for studying concept learning is far from ideal for assessing knowledge effects. Historically, most studies of concept learning share two general characteristics. First, subjects are instructed to learn a concept that distinguishes the members of one category from those of other categories. Second, the studies primarily use linguistically described feature
lists or, sometimes, simple perceptual stimuli. These stimuli usually consist of a small set of unambiguous features.

Both of these characteristics have their virtues. Giving subjects instructions to learn concepts that distinguish one category from another focuses them on an important role of concepts. Namely, concepts allow people to identify novel members of a category. Because the studies involve simple, unambiguous stimuli, both experimenter and subject often agree on the features that characterize them. Typically, this agreement is an essential prerequisite for manipulating variables that may affect learning and for evaluating models of concept learning. To take a simple case, suppose that a researcher hypothesized that conjunctive rules were easier to learn than disjunctive rules, and created artificial categories that could be described by these kinds of rules (e.g., see Medin et al., 1987). To test this hypothesis, subjects necessarily must view the categories as having the features that make up these rules. Furthermore, agreement about features also lets researchers compare directly the concepts formed by computational models with those formed by people. In this case, researchers can be confident that the features that they present to their models will be the same ones that people would consider in forming their concepts. Finally, much of the debate among empirical learning models concerns how features are integrated or combined (e.g., Nosofsky, 1991); without some common ground of features the debate is irrelevant.

On the other hand, this common procedure is not ideal for investigating how prior knowledge affects category learning. Most importantly, a crucial problem in learning is determining the features over which learning takes place (Wisniewski & Medin, 1991, 1994). In many contexts, prior knowledge probably helps people to determine these features. For example, consider the interpretation of x-rays in medical diagnosis. Whether a white area on an x-ray indicates a lung tumor, a bone, or an artifact of the procedure can depend on a variety of sources of prior knowledge, including the patient’s case history, whether the x-ray is a chest x-ray, the physician’s knowledge of the location of bones, and the likelihood that the x-ray was underexposed (see Lesgold et al., 1988, for examples). As another example, in learning about CD players, it is probably the case that not all the features are immediately obvious. Prior knowledge about stereo players and knowledge gleaned from instruction manuals and from other owners may assist one to determine the features.

To summarize, the role of prior knowledge in determining units of analysis or features has not been investigated in traditional learning studies. Typically, features are clearly specified and unambiguous and the instructions encourage subjects to find some set of features with which they can reliably identify the members of a category. Given stimuli composed of a small set of unambiguous features, it is probably not too difficult to find
features that discriminate the members of one category from those of other categories. Prior knowledge is not needed to determine the features.

For these reasons, studying the effects of prior knowledge requires a different experimental paradigm. In our studies, subjects learned about categories of relatively complex stimuli that were given meaningful labels. The category items for our studies were children's drawings of people, shown in Figures 2, 3, and 4. We used labels that were likely to activate prior knowledge in the form of people's intuitive theories about things in the world. For example, some of the drawings of people were labeled as "done by creative children." This label might activate a simple theory such as "creative children are likely to draw better and to depict more difficult aspects of a drawing such as profile and space in three dimensions." Here, the term theory refers to a body of knowledge (often causally connected) that may include scientific principles, stereotypes, and informal observations of past experiences (see Keil, 1989; G. Murphy & Medin, 1985; Nisbett & Ross, 1980). Combining theory-activating labels with complex stimuli creates a situation that is likely to reveal the effects of prior knowledge.

The following studies used a rule-learning paradigm. The general idea is to set up a situation in which there are many potential rules and then to examine the influence of prior knowledge on rule development. As we shall see, manipulating prior knowledge by simply varying category labels has a profound influence on rules. This influence is not readily captured by the standard model augmented with an independent knowledge-driven module that selects or adds features.

The first experiment used a nonincremental rule-learning task in which subjects determined a single rule (or concept) by examining a number of preclassified items that were presented at the same time. This procedure provided a simply way of comparing rules formed by subjects who were operating with or without prior knowledge (i.e., who were given meaningful or neutral labels). The second experiment used an incremental rule-learning task in which unclassified training items were presented one at a time and subjects determined a rule after seeing each item. With this task, we could more carefully examine how people develop and modify their rules as a function of both theories and increasing experience.

The goal of these experiments is to discover some of the ways that prior knowledge and experience interact during learning and to suggest ways that these effects can be incorporated into a model of category learning. The studies also go beyond previous ones that have investigated knowledge effects in category learning. Past studies have primarily demonstrated the facilitatory or inhibitory effects of knowledge. In general, these studies are consistent with the view that knowledge selects or makes relevant certain features during learning. We begin by examining the plausibility of viewing knowledge as a selection process.
Experiment 1a involved a nonincremental rule-learning task in which subjects determined a classification rule by examining categories of items. For one group of subjects (the theory group), the categories were given meaningful labels. It was hypothesized that these labels would activate prior knowledge or expectations about the categories. A second group (the standard group) was given neutral labels and therefore was not hypothesized to have expectations about the category. After examining the categories, subjects classified novel test items. Corresponding to the theory and standard groups were control groups that were not required to produce a classification rule. As described in the following, we systematically varied characteristics of the learning and test items in order to highlight differences in rule learning, theory use, and classification.

The study had a variety of purposes. Importantly, it examined the validity of a knowledge-based selection model of categorization. According to the selection view, both the theory and standard groups should use the same learning process to construct rules, except that the process may operate over different sets of features in the two groups. If prior knowledge only functions to select features, then the only potential difference between the theory and standard groups during learning is the features that describe the category items. Specifically, the theory group may attend to features that prior knowledge has indicated are relevant to the category. The standard group will not be biased in this manner. Therefore, we would expect the rules of the theory and standard groups to have the same basic structure, though they might differ in content. Based on prior studies of human rule induction, one would expect the rules to take the form of either a conjunction of features, a disjunction of features, or a disjunction of conjunctions of features (e.g., Medin et al., 1987). The selection view predicts that both groups will produce rules of this form, but that these rules may involve different features.

We designed the training items so that they were likely to highlight differences in rule learning and theory specialization. There were two important aspects of this selection. First, we explicitly constructed a pair of categories that could be described by both theory-relevant, abstract features and by simple, concrete features. Second, each of these two categories could be described by a different abstract feature that was relevant to the same theory. Specifically, one category contained detailed drawings and the other category contained unusual drawings. Within the theory groups, we counterbalanced the labeling so that for half of the theory group, the detailed category was drawn by creative children, and for the other half, the unusual category was drawn by creative children. As a result, we could compare rules that were produced by the same theories operating on different data (i.e., different categories).
Similarly, we designed the test items so that the expected differences in rule learning and theory specialization would lead to corresponding differences in the classification of the test items. We systematically varied the concrete and theory-relevant features in the test items in such a way that if the theory and standard groups focused on different features, then they would classify the items in different, predictable ways. In addition, we varied whether or not a test item was detailed or unusual so that if people were specializing theories in different ways, they should also classify the test items in different, predictable ways.

Finally, we included control groups that were not asked to give rules in order to assess whether people's rules did in fact correspond to the processes that they used to determine category membership. A number of studies have shown clear discrepancies between the reasons that people give for their behavior and the factors that actually affect their behavior (see Nisbett & Wilson, 1977, for a review). Including novel test items is one way to assess the relation between stated rules and classification judgments. If people's beliefs did correspond to their rules, then we would expect their classification of test items to be consistent with those rules. Even with this correspondence, however, people may have been trying to be consistent with what they say and do, rather than with what they believe and do. Requiring subjects to verbalize their reasons for category membership might alter their actual beliefs about the category or change the learning process in unanticipated ways. Including control groups who were not asked to verbalize rules allows us to examine these possibilities.

Method

Subjects
The subjects were 124 undergraduate students (male and female), attending the University of Michigan, who received course credit for participating in the study.

Materials
The stimuli consisted of drawings (taken from Koppitz, 1984, and Harris, 1963) that were produced by children who were administered the "draw-a-person test." In this task, children are instructed to draw "one whole person." The test is one tool used in psychodiagnosis and IQ assessment. The test is a fairly reliable shorthand indicator of emotional problems and of intelligence in young children (Goodenough & Harris, 1950; Harris, 1963; Koppitz, 1984). On the other hand, its validity as a diagnostic indicator of mental illness in adults has been severely challenged (e.g., L. Chapman & J. Chapman, 1967).

The training items consisted of two categories of six drawings. They were designed so that the drawings in the first category were detailed and those in
the second category were unusual. By coincidence, most of the drawings in the second category also depicted people performing actions. According to pilot studies, people expect all three of these abstract features to be more diagnostic of drawings done by creative children than of those done by non-creative children. (It is unknown whether any of the drawings were actually done by creative children.)

The drawings were also chosen so that at least three simple rules involving concrete features applied to one category and not to the other (some of the drawings were doctored to meet these constraints). An example of one rule was "straight hair or arms at side," which applied to the first category. The drawings were also altered to make the contrast between "detailed" or "unusual" more salient and to assure that the concrete rules applied only to the appropriate category. The authors and a research assistant agreed that the drawings met these constraints. Figure 2 shows the two categories of drawings and the concrete rules that apply to them.

The test stimuli consisted of 16 children's drawings, shown in Figure 3. Conceptually, they can be divided into four sets of 4 drawings, on the basis of whether they are detailed or not detailed and unusual or not unusual. In addition, for each set, 2 of the 4 drawings can be described by the concrete rules that apply to the unusual and active category (see Figure 2). The other 2 drawings can be described by the concrete rules that apply to the detailed category (also shown in Figure 2). For example, one set of drawings shown in Figure 3 (1, 2, 3, and 4) are not detailed and not unusual. Two of these drawings (1 and 2) can be described by the three concrete rules of the unusual category (i.e., "straight hair or arms at the side," "dark-colored shoes and not smiling," "ears and short sleeves"). The other two drawings (3 and 4) can be described by the three concrete rules of the detailed category (i.e., "curly hair and arms not at the side," "light-colored shoes or smiling," "wearing a collar or a tie"). The authors and a research assistant all agreed that the drawings met these constraints.

**Procedure**

About half \((n = 63)\) of the subjects were randomly assigned to the rule group. These subjects were subdivided into those who were told that the drawings were done by "creative and noncreative children" \((n = 32)\) and those who were told that the drawings were done by "children in Group 1 or Group 2" \((n = 31)\). The instructions for those subjects given the theory labels were:

In this experiment, you will be shown two groups of children's drawings. One group was done by Creative children. The other group was done by Non-Creative children. Your task is to come up with a rule that someone could use to decide whether a new drawing belongs to the group drawn by the Creative children. In writing down a rule, it is important that the rule "works" for all
of the drawings in the Creative group and none of drawings in the Noncreative group. That is, if someone were given your rule and the drawings from the two groups (all mixed up), they should be able to use it to divide the drawings into those that belong in one group and those that belong in the other. Also, please write your rule clearly and describe the rule so that the experimenter will be able to understand what you mean. (We will be looking at the rules at a later time). Also, it may be difficult to come up with a rule. Just do the best that you can. After you have finished writing down your rule, the experimenter will hand you a new set of drawings. Examine each drawing and decide whether it was done by a Creative child or a NonCreative child. Indicate your decision (in writing) below each drawing. Thank you.

These instructions were almost identical for those subjects given standard labels, except that "creative" was replaced by "Group 1" and "non-creative" was replaced by "Group 2."

The remaining 61 subjects were assigned to be no-rule group. Like the rule group, they were subdivided into those who were told that the drawings were done by "Creative and NonCreative children" (n = 30) and those who were told that the drawings were done by "children in Group 1 or Group 2" (n = 31). The instructions for those subjects given the theory labels were:

In this experiment, you will be shown two groups of children's drawings. One group was done by Creative children. The other group was done by NonCreative children. Your task is to examine the drawings for 5 minutes and figure out why the drawings belong in the two different groups. It may be difficult to figure out why the drawings belong in the groups that they do. Just do the best that you can. After you have finished looking at the drawings, the experimenter will hand you a new set of drawings. Examine each drawing and decide whether it was done by a Creative child or a NonCreative child. Indicate your decision (in writing) below each drawing. Thank you.

These instructions were almost identical for those subjects given standard labels, except that "Creative" was replaced by "Group 1" and "NonCreative" was replaced by "Group 2."

Categories and their labels were counterbalanced. For about half (n = 30) of the rule/theory and no-rule/theory subjects, the creative drawings were the detailed set and the noncreative drawings were the unusual set. For the remaining (n = 32) subjects, this labeling was reversed. Likewise, for about half (n = 32) of the rule/standard and no-rule/standard groups, the Group 1 drawings were the detailed set and the Group 2 drawings were the unusual set. For the remaining (n = 30) subjects, this labeling was reversed.

Given the procedure described, subjects could belong to one of eight conceptually distinct groups. Subjects were either instructed or not instructed to give a rule, the two categories of drawings were either given theory or standard labels, and the labels were paired with the categories in one of two possible ways.
Detailed Category

Concrete rules:

"curly hair and arms not at the side"
"light-colored shoes or smiling"
"wearing a collar or tie"
Unusual Category

Concrete rules:

“straight hair or arms at the side”
“dark-colored shoes and not smiling”
“ears and short sleeves”

Figure 2. Drawings given the meaningful labels “drawn by creative/noncreative children” (Experiment 1a).
not Detailed & not Unusual

Concrete rules of Unusual and Action Category

Concrete rules of Detailed Category
Detailed & Unusual

Concrete rules of Unusual and Action Category

Concrete rules of Detailed Category

(continued)
Detailed & not Unusual

Concrete rules of Detailed Category

Concrete rules of Unusual and Action Category
Concrete rules of Unusual and Action Category

Concrete rules of Detailed Category

Figure 3. Test items used in Experiment 1a.
Results
The rules of the standard and theory groups differed dramatically in their content and structure. To clarify these differences, we developed a classification scheme for the content and structure of the rules.

Concrete versus Abstract Features
In general, the features mentioned in the rules could be divided into two broad types. *Concrete* features consisted of simple terms that were easily observable in the drawings. They typically referred to parts of the body or to items of clothing (e.g., "arms at sides," "buttons," "pockets," "collar," "elbows bent," "curly hair"). *Abstract* features were more complex, more high-level, or less perceptual (e.g., "action," "true to life," "more depth," "many more details," "positive emotional expression," "bodily expression").

Features Used to Support Other Features
In addition, subjects appeared to use features in several different ways. First, as would be expected, their rules consisted of features that directly determined category membership. Second, subjects sometimes provided features as examples of, or support for, a more abstract feature. We refer to features that have been used in this manner as *support* features. To illustrate these different uses, consider part of the rule that a subject gave for the detailed category, shown in Figure 2: "...much more attention was given to the clothing. There are many more details, things like belts, pockets, and patterns on the clothes..." This rule implies that for a drawing to belong in this category, it must have "details in the clothing." It is having this feature that directly determines category membership. On the other hand, the features "belt," "pocket," and "patterns on the clothes," only indirectly determine category membership. They provide support for the abstract feature "details in the clothing" in the sense that they are examples of, or evidence for, this feature.

Features of Theory versus Standard Group
To examine possible differences in the rules used by the standard and theory groups, we independently classified each feature as either *concrete* or *abstract* and determined whether or not it was used to support a more abstract feature. Overall, we agreed on 85% of their judgments. Some of the differences in scoring were due to simple errors (e.g., an author scoring a feature as "concrete" when he had previously scored occurrences of the feature as "abstract," or an author neglecting to classify a feature). Other differences were due to different interpretations of what the subject meant in a rule and to different intuitions as to whether a feature was abstract or concrete. These differences were readily resolved through discussion.
TABLE 2
Percentages of Rule Types Used by the Theory and Standard Groups (Experiment 1a)

<table>
<thead>
<tr>
<th></th>
<th>Theory (%)</th>
<th>Standard (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concrete</td>
<td>12</td>
<td>49</td>
</tr>
<tr>
<td>Abstract</td>
<td>25</td>
<td>29</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>63</td>
<td>22</td>
</tr>
</tbody>
</table>

Examples of rules used by theory and standard groups for the categories shown in Figure 2 (Experiment 1a).

Concrete
"...all of the characters have their arms out straight from their bodies and they're also standing very straight, facing the front" (for the detailed category).
"Everyone wearing short sleeves, crew-neck style tops" (for the unusual category).

Abstract
"...make drawings that show more positive emotional expression..." (for the detailed category).
"... have angry or dissatisfied looks on their faces" (for the unusual category).

Hierarchical
"...much more attention was given to the clothing. There are many more details in each of them, things like belts, pockets, and patterns on the clothes...more detail in the faces, the noses were more defined as well as the eyes" (for the detailed category).
"...all imply something is being done (an activity, e.g., walking, reading) or was done or is going to be done. They all involve some form of movement" (for the unusual category).

Given these judgments, we classified each rule as hierarchical, concrete, or abstract. A rule was classified as hierarchical if it contained any support features. Otherwise, it was classified as concrete if it contained a majority of concrete features, or as abstract if it contained a majority of abstract features.

There were several striking differences in the kinds of rules used by the standard and theory groups. Table 2 shows the percentages of hierarchical, concrete, and abstract rules for each group, and provides examples of each type of rule for the categories shown in Figure 2. The standard group primarily used concrete rules and used abstract and hierarchical rules to a lesser extent. Specifically, 48% of the standard group's rules were concrete, 29% were abstract, and 22% were hierarchical. Even though simple rules of concrete features were possible, the theory group primarily used abstract and hierarchical rules (which consisted of abstract features and other features used to support them). Specifically, 25% of their rules were abstract, 63% were hierarchical, and 12% were concrete.
These differences were significant. The theory group used more abstract features per rule than the standard group (2.58 vs. 1.58), $t(62) = 2.28, p < .05$. In addition, the theory group used more support features per rule than the standard group (2.59 vs. 1.02), $t(62) = 2.76, p < .01$. On the other hand, the standard group used more concrete, nonsupporting features per rule than the theory group (2.00 vs. .99), $t(62) = 2.54, p < .02$.

There were also clear differences in the use of theory-relevant features by the groups. When the detailed category was labeled "creative," 13 of 16 subjects (81%) mentioned some aspect of detail in their rules about creative children. When the same category was labeled "Group 1," only 5 of 16 subjects (31%) mentioned some aspect of detail in their rules about Group 1. This difference was significant, $\chi^2 (1, N=32) = 8.14, p < .005$. When the unusual (and action) category was labeled "creative," 11 of 16 subjects (69%) mentioned some aspect of action in their rules about creative children. (No one mentioned that these drawings were unusual.) When the same category was labeled "Group 1," only 4 of 15 subjects (27%) mentioned some aspect of action in their rules about Group 1. This difference was also significant, $\chi^2 (1, N=31) = 5.38, p < .02$.

**Classification of Test Items**

Note that classifying a test item implicitly corresponds to placing it into one of the two categories previously seen during learning. Table 3 lists the 16 test items and the percentage of subjects in each of the eight groups who placed a given item in the detailed category. (The percentages in parentheses are those for the theory and standard groups who were not asked to give rules.) For those subjects who had seen the detailed category labeled as "creative" ("noncreative") during learning, these percentages reflect the proportion of subjects who classified a test item as "creative" (or "non-creative"). In the figure, the left-most column lists each test item by number and indicates whether it was detailed or not detailed, unusual or not unusual, and whether it could be described by the concrete rules of the detailed category or those of the unusual category (the actual test items are shown in Figure 3). To understand the table better, consider the percentages listed in the second column. In column 2, each percentage reflects the proportion of subjects who, during learning, had seen the detailed category labeled "creative" and who, during testing, classified a given test item as "creative," thus placing it into the detailed category. Notice, for example, that Drawings 1 through 4 are not detailed, and as would be expected, the theory/rule and theory/no-rule groups did not tend to place these drawings into the detailed category (i.e., classify them as creative). On the other hand, Drawings 5 through 8 are detailed, and these groups did tend to call these drawings "creative," thus placing them in the detailed category.
<table>
<thead>
<tr>
<th>Test items</th>
<th>Column 1</th>
<th>Column 2</th>
<th>Column 3</th>
<th>Column 4</th>
<th>Column 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 — detailed, — unusual concrete rules=unusual</td>
<td>13 (14)</td>
<td>6 (19)</td>
<td>50 (44)</td>
<td>67 (60)</td>
<td></td>
</tr>
<tr>
<td>2 — detailed, — unusual concrete rules=unusual</td>
<td>13 (29)</td>
<td>69 (69)</td>
<td>13 (44)</td>
<td>73 (60)</td>
<td></td>
</tr>
<tr>
<td>3 — detailed, — unusual concrete rules=detailed</td>
<td>25 (21)</td>
<td>25 (13)</td>
<td>75 (50)</td>
<td>33 (53)</td>
<td></td>
</tr>
<tr>
<td>4 — detailed, — unusual concrete rules=unusual</td>
<td>38 (36)</td>
<td>0 (13)</td>
<td>69 (88)</td>
<td>33 (33)</td>
<td></td>
</tr>
<tr>
<td>5 detailed, unusual concrete rules=unusual</td>
<td>88 (71)</td>
<td>94 (94)</td>
<td>19 (25)</td>
<td>73 (67)</td>
<td></td>
</tr>
<tr>
<td>6 detailed, unusual concrete rules=unusual</td>
<td>69 (57)</td>
<td>81 (81)</td>
<td>25 (13)</td>
<td>93 (53)</td>
<td></td>
</tr>
<tr>
<td>7 detailed, unusual concrete rules=detailed</td>
<td>81 (64)</td>
<td>94 (94)</td>
<td>25 (25)</td>
<td>53 (67)</td>
<td></td>
</tr>
<tr>
<td>8 detailed, unusual concrete rules=detailed</td>
<td>75 (71)</td>
<td>81 (94)</td>
<td>25 (19)</td>
<td>53 (53)</td>
<td></td>
</tr>
<tr>
<td>9 detailed, — unusual concrete rules=detailed</td>
<td>88 (71)</td>
<td>38 (19)</td>
<td>88 (75)</td>
<td>47 (47)</td>
<td></td>
</tr>
<tr>
<td>10 detailed, — unusual concrete rules=detailed</td>
<td>81 (86)</td>
<td>81 (88)</td>
<td>44 (69)</td>
<td>47 (33)</td>
<td></td>
</tr>
<tr>
<td>11 detailed, — unusual concrete rules=unusual</td>
<td>81 (86)</td>
<td>88 (75)</td>
<td>38 (63)</td>
<td>80 (53)</td>
<td></td>
</tr>
<tr>
<td>12 detailed, — unusual concrete rules=unusual</td>
<td>75 (57)</td>
<td>88 (69)</td>
<td>31 (56)</td>
<td>67 (53)</td>
<td></td>
</tr>
<tr>
<td>13 — detailed, unusual concrete rules=unusual</td>
<td>6 (14)</td>
<td>25 (31)</td>
<td>44 (25)</td>
<td>73 (58)</td>
<td></td>
</tr>
<tr>
<td>14 — detailed, unusual concrete rules=unusual</td>
<td>63 (79)</td>
<td>94 (100)</td>
<td>6 (19)</td>
<td>73 (60)</td>
<td></td>
</tr>
<tr>
<td>15 — detailed, unusual concrete rules=detailed</td>
<td>31 (29)</td>
<td>0 (6)</td>
<td>75 (56)</td>
<td>27 (33)</td>
<td></td>
</tr>
<tr>
<td>16 — detailed, unusual concrete rules=detailed</td>
<td>44 (21)</td>
<td>50 (31)</td>
<td>56 (81)</td>
<td>67 (33)</td>
<td></td>
</tr>
</tbody>
</table>
Classification of Test Items: Theory versus Standard Groups

The theory and standard groups classified the drawings in dramatically different ways. We compared the classification percentages (shown in Table 3) of the theory and standard groups, which differed only in whether they were given meaningful or neutral category labels. The overall correlation between the classification percentages of these groups was only .12 (n.s.).

Importantly, the presence or absence of the theory-relevant features in the drawings affected the classifications of the theory group. On the other hand, the presence or absence of the concrete features affected the classifications of the standard group. To show this, we examined the classification of test items that would have been placed into different categories if the theory group had classified them using the theory-relevant features and if the standard group had classified them using the concrete features.

First, consider those subjects in the theory group who were told that the detailed category consisted of “drawings done by creative children” versus those subjects in standard group who were told that this category consisted of “drawings done by children in Group 1.” If these groups attended to the theory-relevant feature “detail,” then they should classify the Test Items 5, 6, 11, and 12, as members of the detailed category and Items 3, 4, 15, and 16 as members of the nondetailed category (see Figure 3). On the other hand, if these groups attended to any of the concrete rules that described the detailed category, then they should classify test items in an opposite manner. That is, Drawings 5, 6, 11, and 12 should be classified as members of the nondetailed category and Drawings 3, 4, 15, and 16 should be classified as members of the detailed category. (Both groups should categorize the other 8 drawings into the same categories, so these are not relevant to the analysis.)

Second, consider those subjects in the theory group who were told that the unusual category consisted of “drawings done by creative children” versus those subjects in the standard group who were told that this category consisted of “drawings done by children in Group 1.” Originally, the test items were designed to be unusual or not unusual. However, because the theory group attended to “action” and not to “unusual,” we reassessed whether the test items showed or did not show action. Two graduate students who had no prior knowledge about this study judged the drawings as showing or not showing action. They judged Test Items 5, 6, 7, 8, 9, 10, 12, and 14 as showing action, and Test Items 1, 2, 3, 4, 11, 13, 15, and 16 as now showing action. If these groups attended to the theory-relevant feature “action,” then they should classify the Test Items 7, 8, 9, and 10 as members of the action category and Items 1, 2, 11, and 13 as members of the nonaction category (see Figure 3). On the other hand, if these groups attended to any of the concrete rules that described the action category, then they should classify test items in an opposite manner. That is, drawings 7, 8,
9, and 10 should be classified as members of the nonaction category and drawings 1, 2, 11, and 13 should be classified as members of the action category. (Both groups should categorize the other 8 drawings into the same categories, so these are not relevant to the analysis.)

Given these predicted patterns of classification, the theory group classified 67% of the preceding test items in a manner consistent with the use of the theory-relevant features. As predicted, the standard group showed the opposite pattern. Subjects in this group classified the same test items 65% of the time in a manner consistent with the use of the concrete features. This difference was highly significant, \( t(122) = 6.60, p < .001 \).

**Classification of Test Items: Rule versus No-Rule Groups**

Asking people to give a classification rule had little effect on how they classified test items. To show this, we compared the classification percentages of groups that differed only in terms of whether or not they had formulated a classification rule. Each pair of groups was given the same labels applied to the same categories, but only one of the two groups was required to give a rule. There were four pairs of groups that differed in this manner. As an example, consider the second column in Table 3. This column lists the classification percentages for the theory group who saw that the detailed (unusual) set of drawings labeled as "creative" ("noncreative") and who gave a rule for category membership. This column also lists (in parentheses) the percentages of the corresponding theory group who did not give a rule for category membership. The third, fourth, and fifth columns of Table 3, respectively, correspond to the classification percentages of the other pairs of rule and no-rule groups. The correlation between the classification percentages of these four pairs of groups was .84 \((p < .001)\). Furthermore, there were few reliable differences between the classifications of items by a rule group and its corresponding no-rule group. Because there were four pairs of such groups and 16 test items, there were 64 possible comparisons. Using a Pearson chi-square test of association, only 2 of the 64 comparisons were reliably different \((p < .05)\). The average chi-square value for these comparisons was only .953. The lack of an effect of the rule/no-rule manipulation contrasts with the strong labeling effect.

**Differences in Theory Specialization**

Recall that the test items varied in terms of whether or not they were detailed or unusual. By way of this manipulation, we predicted that one theory group would label test items "creative" and that the other theory group would label "noncreative" (and vice versa). This prediction follows from the view that people's theories can be specialized in different ways. In this study, both theory groups presumably had similar prior expectations about creativity (i.e., that they will be detailed, unusual, action-oriented, etc.).
However, because the groups viewed different examples of drawings done by creative children, their theories may have been specialized differently. As a result of learning, one group may have arrived at the belief that creative drawings show detail. The other group may have arrived at the belief that creative drawings are unusual. We would then have expected one theory group to label Drawings 9 through 12 as "creative" (because they were detailed) and Drawings 13 through 16 as "noncreative" (because they lacked details). Conversely, we would have expected the other theory group to label in an opposite way: Drawings 9 through 12 would be labeled as "noncreative" (because they are not unusual) and Drawings 13 through 16 would be labeled "creative" (because they are unusual).

As mentioned, subjects who were told that the unusual category was the creative one, generally interpreted those drawings as showing action (rather than being unusual). Unfortunately, whether a test item showed or did not show action was almost perfectly correlated with whether it showed or did not show detail. For 14 of the 16 test items, a drawing that was detailed also showed action, or a drawing that was not detailed did not show action. As a result, it was generally not possible to evaluate differences in theory specialization by examining the classification data.

On the other hand, the rules of the two theory groups clearly showed that they specialized their theories in different ways. As noted, when the detailed category was labeled "creative," 13 of 16 subjects mentioned some aspect of detail in their rules about creative children. Only 1 subject mentioned detail as evidence of creativity when the unusual (and action) category was labeled "creative." This difference was highly significant, $\chi^2 (1, N=32) = 18.29, p < .001$. Conversely, when the unusual (and action) category was labeled "creative," 11 of 16 subjects mentioned some aspect of action in their rules. No subject mentioned action when the detailed category was labeled as "creative." This difference was also highly significant, $\chi^2 (1, N=32) = 16.76, p < .001$.

Discussion

Our results are damaging to the knowledge-as-feature-selection position. The selection view implies that the rules of the theory and standard groups should have the same form or structure, although they might contain different features. Specifically, the theory group might be more likely to use knowledge-relevant features in their rules.

The results did not conform to these predictions. First, there were structural differences between the rules of the two groups. The standard group tended to form rules that were flat. That is, their rules consisted primarily of concrete features that were roughly at the same level of abstraction (see Table 2 for examples). On the other hand, a fair number of rules of the theory group were hierarchical. That is, they contained features at lower
levels of abstraction (i.e., concrete features) linked to, or used to support, those at higher levels (i.e., abstract features; see Table 2 for examples).

Importantly, these structural differences suggest that knowledge may actually affect the learning algorithm that people use (see also Wattenmaker et al., 1986). The standard and theory groups appeared to use different algorithms in forming their rules. The strategies and rules of the standard group were similar to those of subjects in previous studies of rule learning that did not involve prior knowledge (e.g., Medin et al., 1987). They are generally consistent with various versions of the standard model (e.g., INDUCE and PATCH; Medin et al., 1987). In particular, the standard group focused on finding a conjunction or disjunction of simple, concrete features (or some combination of conjunctions and disjunctions of these features) that applied to one category and not to the other.

In contrast, the theory group seldom used the learning strategies described previously. Learning did not conform to the standard model, nor to the standard model augmented with a knowledge-driven selection process. Instead, the theory group appeared to search for evidence (i.e., concrete features) in the data that supported abstract features and hypotheses. As a result, the theory group often formulated hierarchical rules, consisting of features at multiple levels of abstraction. The abstract features were probably activated by people's intuitive theories about the children who drew the pictures. For example, consider the first hierarchical rule shown in Table 2. In this case, the meaningful label, "drawn by creative children" may have activated the abstract hypothesis that creative children draw better and are more observant. This hypothesis, in turn, might suggest that the drawings of creative children will show details such as belts, pockets, patterns on the clothes, and more defined noses. In short, rule learning for the theory groups seems to be as much a vertical as a horizontal process and flat features will be inadequate.

The results of this study also suggest that our claims are not an artifact of asking people to give linguistic descriptions of their thought processes. The pattern of classification of novel items was very similar for subjects who differed only in terms of whether or not they had produced a rule during learning. In contrast, the pattern of classification was very different for subjects who were given different labels for the categories.

The study also demonstrates that people specialize similar theories in different ways, when different data are coupled with those theories. Recall that the theory groups tended to believe that creative children would draw detailed pictures or that they would draw pictures depicting action. They specialized these abstract features accordingly. However, the particular abstract feature that they specialized (i.e., "detail" or "action") was strongly determined by the data (i.e., whether the label "drawn by creative children" was paired with the detailed category or the unusual and action category). Our
interpretation of this finding is that subjects did not have strong, specific theories about the children’s drawings. Rather, their theories probably consisted of a cluster of general beliefs or biases about drawings done by creative children, derived from commonsense notions of creativity, children, and what people look like. The data in turn, were relatively free to “manipulate” these beliefs.

**EXPERIMENT 1b**

The next experiment attempted to replicate the major finding of Experiment 1a that theories activate abstract features and learning involves searching for evidence in the data that provides support for these features. This finding suggests that prior knowledge does not solely involve the selection of relevant features. The experiment went beyond the previous one in that it involved more categories of drawings and a greater variety of theories.

**Method**

**Subjects**
The subjects were 40 undergraduates (male and female), attending the University of Michigan, who received course credit for participating in the study.

**Materials**
The stimuli were three sets of 10 drawings, divided into two equal-size categories. One of these sets is shown in Figure 4.

**Procedure**
A subject was randomly assigned to the theory group or to the standard group. The theory group viewed three sets of categories that were given meaningful labels (see the following). The standard group viewed the same sets of categories (and some other sets that were not included in the analyses described later). Their drawings however, were simply labeled “Group 1” and “Group 2.” The order that subjects viewed the sets was counterbalanced: approximately an equal number of subjects viewed each set before the other two. The instructions for the theory and standard groups were identical to those of the theory and standard groups of Experiment 1a.

For the theory groups, we used different meaningful labels to explore a variety of simple theories and expectations that might be activated. Specifically, one set of categories was labeled, “drawn by children with high IQs/low IQs,” a second set was labeled, “drawn by creative/noncreative children,” and a third set was labeled, “drawn by emotionally disturbed/well-adjusted children.” The categories were arbitrarily selected groups of children’s drawings with the exception of two categories in one set that were actually done by emotionally disturbed and well-adjusted children, respectively.
Labeling of the categories was counterbalanced in the theory group. So, for example, half of the theory group was told that Category 1 (shown in Figure 4), was “drawn by low-IQ children” and that Category 2 was “drawn by high-IQ children.” For the other half of the group, this labeling was reversed. The counterbalancing was done in order to explore the effects of the same theories applied to different data (i.e., when identical labels were given to nonidentical categories) and different theories applied to the same data (i.e., when different labels were given to identical categories).

Results

As in the previous experiment, the rules consisted of features that were either concrete or abstract and features that were used to support more abstract features. We classified the features into these three categories, agreeing on 87% of their judgments. These differences were readily resolved through discussion. Given these judgments, rules were again classified as hierarchical, concrete, or abstract.

In general, the pattern of rule use was similar to that found in Experiment 1a. The standard group primarily used concrete rules and used abstract and hierarchical rules to a lesser extent. Unlike the first study, however, they sometimes used patched rules (which also consisted primarily of concrete features). Specifically, 71% of the rules of the standard group were concrete, 22% were abstract, and only 7% were hierarchical. On the other hand, the theory group primarily used abstract rules or hierarchical rules, although they produced a small proportion of hierarchical rules than in Experiment 1a. Specifically, 38% of the rules of the theory group were abstract, 34% were hierarchical, and 28% were concrete.

These differences were significant. The theory group used more abstract features per rule than the standard group (1.75 vs. .75); this difference was highly significant, $t(38) = 4.19, p < .001$. In addition, the theory group used more support features per run than the standard group (.74 vs. .09); this difference was also highly significant, $t(38) = 3.27, p < .01$. Most (85%) of the theory group used support features in at least one of their rules, compared to only 30% of the standard group. On the other hand, the standard group used more concrete (nonsupporting) features per rule than the theory group (2.36 vs. .75); this difference was highly significant, $t(38) = 4.86, p < .001$.

Effects of Counterbalancing in the Theory Groups

Recall that the meaningful category labels were counterbalanced in order to compare rules produced by different theories operating on the same data (i.e., rules for identical categories given different labels) and to compare rules that were produced by similar theories operating on different data (i.e., rules for nonidentical categories given the same labels).

In general, it was not possible to compare rules produced by different theories applied to the same data (i.e., same category). With two of the sets
Category 1

1

2

3

4

5
Figure 4. Drawings given the meaningful labels "drawn by low-IQ/high-IQ children" (Experiment 1a) and "drawn by creative/noncreative children" (Experiment 1b).
of labels (high IQ vs. low IQ and creative vs. noncreative), subjects tended only to produce rules that applied to the categories labeled “high IQ” and “creative” and seldom produced rules for the categories labeled “low IQ” and “noncreative.” As a result, we could not compare the rules (for example) of the subjects who were told that Category 1 of Figure 4 was “drawn by high-IQ children” to the rules of subjects who were told that this same category was “drawn by low-IQ children.”

On the other hand, as in Experiment 1a, there was evidence that similar theories, when applied to nonidentical categories, yielded different rules. Such differences were particularly apparent for Category 1 and Category 2 that were given the labels “drawn by emotionally disturbed/well-adjusted children.” Specifically, subjects mentioned “lack of detail” or “simplicity” in four of eight rules that applied to Category 1 when it was labeled “drawn by emotionally disturbed children,” but the other subjects failed to mention these features in any rules that applied to the identically labeled Category 2. Instead, subjects mentioned that the drawings of this category 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 that applied to the identically labeled Category 1.

Discussion

Taken together, Experiments 1a and 1b undermine the knowledge-as-selection view of learning. One cannot assume that prior knowledge only functions to select features from a preexisting set and that some version of the standard model of learning is then applied (as in Figure 1a). Rather, people with prior knowledge about a category (i.e., the theory groups) appear to approach learning in a qualitatively different way and to use knowledge to infer rather than only to select features. Most notably, prior knowledge activates abstract features and people attempt to specialize them, given the data. The great many more hierarchical rules in the theory groups than in the standard groups suggests that this was the case. Furthermore, the predominance of abstract features in the rules of the theory groups and the lack of such features in the rules of the standard group suggests that these features were added by prior knowledge. Otherwise, it is difficult to explain why the standard group generally failed to use these features in their rules.

Given the insufficiency of the knowledge-as-selection model, one might be tempted to account for the current findings with a knowledge-as-addition model, because we have claimed that knowledge also functions to add features to a training item. Recall that in this model, knowledge selects features and adds features to the item and that some version of the standard model of learning is then applied (see Figure 1a). This view is also incom-
patible with the current findings. These findings strongly suggest that theory-driven learning involves a search for evidence in training items that supports or provides evidence for abstract features. That is, people are trying to link concrete features of the training items to abstract features that, in turn, are linked to the more abstract category labels. In contrast, the knowledge-as-addition model assumes that the learning process is the same as in the selection model: People search for features among training items that describe a category and that distinguish it from other categories. The only difference between the selection and addition models is that learning operates over different feature spaces.

The next study more closely examines how knowledge and experience interact during learning, suggesting that the effects of knowledge and experience are tightly coupled and interwoven. It also provides further kinds of evidence against the selection and addition models. The current methodology has some limitations. First, people seldom acquire rules or concepts by examining a number of preclassified examples at the same time (as they did in these studies). Typically, rule learning is incremental and items are not preclassified. Second, the methodology only enables us to investigate the final output of people's rule-learning processes. We would like to investigate how rules are developed and modified in response to new category items. The next study uses a different methodology that avoids these limitations and focuses more on the process of rule learning. It more carefully examines the effect of experience on prior knowledge, and vice versa.

**EXPERIMENT 2**

Experiment 2 used an incremental rule-learning task in which items were presented one at a time. The task for the subject was to classify each item into one of two categories and to give a rule or reason for the classification. Subjects were then given feedback about their classification decisions and given the opportunity to modify their rules. As before, categories were assigned meaningful labels that were intended to activate prior knowledge. In addition, we asked people to give us a "first-impression" rule of the category, before they had seen any items. We assumed that this rule reflects a person's prior knowledge or expectations about a category before having experience with it. In this way, we could examine how experience (i.e., training items and feedback) in conjunction with people's prior beliefs affect rule development and modification.

Experiment 2 enables us to examine how people modify their rules or concepts about a category in response to feedback. The selection and addition models make some clear predictions about the modification process. To examine these predictions, consider the simple case in which someone is learning about two categories and has formulated a rule for each category.
The person then incorrectly classifies an item. In this context, the most straightforward reason for why the person failed to classify the item correctly was that the rule for the correct category failed to apply and/or the rule for the incorrect category did apply. In other words, the rule for the correct category is too specific (i.e., it does not apply to an item when it should) and/or the rule for the incorrect category is too general (i.e., it applies to an item when it should not).

The standard model (and therefore the selection and addition models) has typically proposed two ways that such rules should be modified. First, an overly general rule may be prevented from applying by adding to the rule one or more features that are true of previously seen category items but are not true of the current, incorrectly classified item. Second, deleting one or more features from the rule may enable an overly specific rule to now apply to the current item. In contrast, the tightly coupled approach to category learning (illustrated in Figure 1a), suggests that learning may involve modifying the features of a rule in addition to adding or deleting features. A major purpose of this study is to compare people's rule modification strategies to these proposals.

The study had a variety of other goals. It attempted to replicate findings of previous studies. It also more closely examined the effects of different theories (activated by different labels) on the same data (i.e., the same group of drawings) and the same theories on different data (i.e., different groups of drawings).

Method

Subjects
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.

Materials
The stimuli were the set of children's drawings shown in Figure 4. Category membership conformed to the partitioning associated with Figure 4 but the labels assigned to these categories varied across groups.

Procedure
Students learned about the categories in the following manner. First, before seeing any drawings, one group (the farm/city group) was told that the categories were “drawing done by farm and city kids.” The other group (the creative/noncreative group) was told that the categories were “drawings done by creative and noncreative kids.” (Unlike the previous studies, all groups were given meaningful labels.) Then, they wrote down their initial, first-impression rules for classifying the drawings. The specific instructions for subjects in the farm/city group were:
I'm going to show you a bunch of drawings done by farm kids and city kids. They are all drawings of kids. Your task is to figure out which drawings were done by farm kids and which ones were done by city kids. I'm especially interested in the kinds of reasons or rules that you use for deciding that a drawing was done by a farm kid or a city kid. Before we actually begin, what are your impressions about the drawings of kids that farm and city kids might do? How do you think the drawings done by farm kids are alike? How do you think drawings done by farm and city kids are different? Please tell me the rule that you would use to decide that a drawing was done by a farm kid and the rule that you would use to decide that a drawing was done by a city kid.

(These instructions and those that follow were identical for the creative/noncreative group except that "farm" and "city" were replaced by "creative" and "noncreative.") The experimenter then wrote down the subject's rules. Next, subjects were shown the drawings one at a time, presented in a random order and had to determine the category in which a drawing belonged. The basic sequence and specific instructions began as follows:

Now I'm going to show you the first drawing. Was this one done by a farm kid or a city kid and why do you think so?

(Experimenter writes down rule and then flips drawing over. On the back of the drawing is the label "farm" or "city.")

What do you think now about drawings done by farm kids versus city kids? That is, after having seen a drawing, what reason or rule would you use for deciding that a new drawing was done by a farm kid or a city kid?

(Experimenter then writes down any modified rule.)

After two drawings were presented in this manner, the experimenter then stated:

Okay, now I'm going to show you some more drawings, one at a time. I'd like you to look at each drawing, decide whether a farm kid or city kid drew it, and write down on this sheet of paper (handing sheet of paper to subject) what rule you used to make your decision. I'll then flip the drawing over and we'll see if you were right. After seeing the drawing, and finding out who drew it, you may wish to change your original rule. Or, you may decide on a new one. Please write the new or changed rule down.

Subjects then viewed 8 more drawings (for a total of 10), receiving feedback after each and having the option of modifying their rules. Afterward, subjects wrote down their overall rule for classifying drawings into the two categories. Here were the specific instructions:

Okay, you've had a chance to study the drawings of farm and city kids. You've had a fair amount of experience in classifying these drawings. Based on your entire experience, I'd like you to give me the best rule for deciding that a drawing was done by a farm kid and the best rule for deciding that a drawing was done by a city kid.
As in previous experiments, labeling of the categories was counterbalanced. Half of the creative/noncreative group was given feedback that the drawings in Category 1 were “done by creative children” and that drawings in Category 2 were “done by noncreative children.” For the other half of the creative/noncreative group, this feedback was reversed. The farm/city group received feedback in an analogous manner.

Results
The results once again demonstrated a major influence of prior knowledge on rule learning. First, subjects used a variety of strategies in modifying their rules, some of which suggested that knowledge and experience interacted to modify features associated with the drawings (as predicted by the tightly coupled approach). Second, the different labels activated different prior expectations about the drawings that, in turn, strongly affected the specific content of the rules and the classification of the drawings. Third, the general content and structure of the rules were similar to that of the theory groups in the first experiment. Like that study, people tended to form hierarchical and abstract rules. In the following, we describe these results in more detail.

Rule-Modification Strategies
Recall that after classifying a drawing and writing down a rule, participants received feedback on their decisions. After feedback, participants had the option of modifying their rules. In general, when people received positive feedback, they usually did not modify their rules (27% of the rules were modified after positive feedback). On the other hand, when given negative feedback, people often modified their rules (71% of the rules were modified after negative feedback). We looked over these responses and derived a set of 7 categories that appeared to capture the common strategies of the subjects. In the following, we describe these response categories and provide examples of those that are not self-explanatory.

1. **Features Support Correct Category.** In this case, people consider features that they had used as evidence for the incorrect category to be evidence for the correct category. For example, one person mentioned the “almost perfect body proportionment” as evidence that a drawing was done by a creative child. Given negative feedback, the person stated “perfect body proportions show lack of imagination,” which is evidence that the drawing was done by a noncreative child. As a second example, a person noted that a drawing was abstract and depicted a person of an ethnic origin, and therefore that it was done by a city child. Given negative feedback, the person stated “it seems to be better if I reverse the requirements.”
2. Other Features Support Correct Category. In this case, people abandon features that they had used as evidence for the incorrect category and shift to other features that provide support 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 non-creative. Told the drawing was creative, the person stated, “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.” (Several participants in the farm/city group believed that plaid clothing was typical of people from rural areas.)

3. Features Are Not Diagnostic. In this case, people note that the features they used to classify a drawing incorrectly are not diagnostic of that category. For example, one person stated a drawing depicted “shoes for outdoor use” and was therefore drawn by a farm child. Upon receiving negative feedback, the person stated “the shoes aren’t a good judge.” Sometimes, people also explained why the features were not diagnostic. For example, one person stated that a drawing was done by a creative child because it “expresses emotion, action—not just a standing person.” Given negative feedback, the person responded, “it is hard to judge through a picture—if the child is a good artist they can convey expression in a special way if they aren’t ‘creative’.”

4. Shift Criterion for Applying Feature. People sometimes 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. However, given negative feedback, the person responded, “drawings done by creative children would be more detailed.” In this example, the person adjusted the belief that creative children will show detail. Specifically, now 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, given negative feedback, the person reexamined the details of the drawing and stated: “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 adjusted or altered the use of the feature detail.
TABLE 4
Rule-Modification Strategies: Number of Responses per Group (Experiment 2)

<table>
<thead>
<tr>
<th></th>
<th>Creative/Noncreative</th>
<th>Farm/City</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Features support correct category</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>2. Other features support correct category</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>3. Features are not diagnostic</td>
<td>16</td>
<td>5</td>
</tr>
<tr>
<td>4. Shift criterion for applying feature</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>5. Reinterpret features</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>6. Same features still distinguish categories</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>7. Uncertain</td>
<td>17</td>
<td>16</td>
</tr>
</tbody>
</table>

Specifically, now for a drawing to be classified as done by a city child, it must show detailed clothing as opposed to detail in body movement.

5. **Reinterpret Features.** Given feedback that they had incorrectly classified a drawing, people sometimes reinterpreted the features that they had used in making the incorrect decision. Then, they used these reinterpreted features as evidence for the correct category. 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, told that the drawing was done by a farm child, the person reinterpreted the character as one created from a farm child’s imagination. 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.”

6. **Some Features Still Distinguish Categories.** In this case, subjects state that their rule still applies (even though they received negative feedback after applying it).

7. **Uncertain.** Subjects state that there is no rule for distinguishing the categories, that they do not know why a particular drawing belongs in the category or that they decide by guessing.

Each author independently classified the subject responses into one or more of these categories (subjects sometimes responded in more than one way). We agreed on 75% of their judgments. Although this agreement is relatively low, a response could be potentially classified into a variety of categories. The remaining differences were resolved by discussion. Table 4 lists the number of occurrences of each response category for the two groups.

Four types of responses occurred more than 10% of the time. When subjects were told that they had incorrectly classified a drawing, they most often mentioned other features that provided support for the correct
category. Subjects responded in this manner 30% of the time. This response is also consistent with the prediction of the selection or addition model that rules are modified by adding features to them. Subjects also mentioned that they were guessing or uncertain about category membership (22% of the responses were of this type). They also mentioned that the features involved in their incorrect classifications were not diagnostic of category membership (14% of the responses were of this type). This response is consistent with the prediction of the selection or addition model that rules are modified by deleting features from them. Finally, subjects reinterpreted features so that they provided support for the correct category (13% of the responses were of this type).

Despite evidence for the modification strategies predicted by the selection or addition model, two of the ways that subjects responded are problematic for these models: reinterpreting features (13% of the responses) and shifting the criteria used for feature interpretation (7% of the responses). Later, we discuss the implications of these findings for a concept-learning system.

**Prior Expectations**

To determine prior expectations, we examined the initial rules of the two groups (given by subjects before any training items had been presented). It is clear that the different labels activated different prior expectations. Most noticeably, a relatively large number of subjects in the creative/noncreative group (13 of 20) explicitly mentioned that the drawings of the creative children would be more detailed or contain more features or shapes than those of the noncreative children. 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." Three other expectations mentioned by a few subjects were that the creative drawings would be more imaginative, unusual, or expressive than the noncreative drawings. On the other hand, most of the subjects in the farm/city group (14 of 20) implied that the drawings would reflect the kinds of people that farm and city kids were exposed to in their different environments and, consequently, that these people could be identified by differences in clothing. 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." Two other expectations mentioned by a few subjects were that the city drawings would be more detailed than the farm drawings and that there would be no difference between the drawings.

**The Effect of Prior Knowledge on the Content of Rules**

The rules that the two groups formed were often related to their prior expectations. As one way to demonstrate this relation, we determined those rules
that were based on the most common prior expectations of the two groups. Each author independently judged whether or not a rule referred to details of a drawing (or lack of) or to a person depicted in the drawing who was likely to come from either a rural or urban environment. Overall, we agreed on 89% of their judgments. We resolved any differences in our scoring through discussion. On the average, a person in the creative/noncreative group used 6.05 rules that referred to details of the drawing compared to 3.05 for a person in the farm/city group. The difference was highly significant, t(38) = 2.94, p < .01. On the other hand, a person in the farm/city group used an average of 6.70 rules that referred to a person depicted in the drawing who was likely to come from a rural or urban environment. In contrast, no one in the creative/noncreative group ever used such a rule. The different prior expectations also caused the groups to use different features in these rules.

There was also evidence that the two groups sometimes interpreted the same part of a drawing differently, and that the different interpretations were related to their different expectations. This finding provides one explanation for why a group used features in their rules that were not used by the other group. To illustrate this finding, consider some of the drawings shown in Figure 4. In Drawing 3, a person in the creative/noncreative group interpreted the vertical line of dots 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 5 depicted “someone dancing.” This feature in turn “showed imagination” (another one of the prior expectations of this group) 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 third example, a person in the creative/noncreative group interpreted the circular configuration of lines in Drawing 2 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 someone from the city and therefore was drawn by a child from the city. As a final example, 5 subjects in the creative/noncreative group interpreted Drawing 8 as a person from a culture different from that of America, for example, a Frenchman (which implied that the drawing was done by a creative child). In contrast, 6 subjects in the
farm/city group interpreted this drawing as a person from a city, for example, a bellboy (which implied that the drawing was done by a city child).

The Effect of Prior Knowledge on Classification
In some cases, subjects' prior expectations strongly affected how they classified the drawings. Table 5 shows the percentage of subjects who classified a given drawing as "creative" (for the creative/noncreative group) and as "city" (for the city/farm group). Note that almost all of the creative/noncreative group classified Drawings 4 and 5 as "creative" and almost all of the farm/city group classified Drawings 1 and 8 as "city." (The binomial probability of 85% choices of one category is less than .01 and for 95% choices, less than .0001). As we suggest later, the classification data for these four drawings are problematic for the addition model, because it has no mechanism for linking knowledge to specific beliefs about category membership.

General Content and Structure of Rules
The rules that subjects gave were similar in structure and in general content to those of the theory groups in the previous experiments. Each author independently classified a rule as hierarchical, concrete, or abstract. A rule was classified as hierarchical if it contained any support features. Otherwise, it was classified as concrete if it contained a majority of concrete features, or as abstract if it contained a majority of abstract features. Overall, we agreed on 87% of their judgments. We resolved any differences in scoring in a similar manner as before. The results generally followed those of the previous experiments: 62% of the rules were abstract, 20% were hierarchical, and 18% were concrete. Within the two groups, these
percentages were: 69% abstract rules, 19% hierarchical rules, and 12% concrete rules for the creative/noncreative group, and 52% abstract rules, 23% hierarchical rules, and 29% concrete rules for the farm/city group.

Discussion
Several results from this experiment present problems for the selection and addition models. To understand why, consider how these models would operate in the context of the incremental rule-learning task. Given a training item, theories or prior expectations are used to infer and/or to select relevant features. After this step, the item is classified according to some version of the standard learning model. As in Figure 1a, for example, a person would compare the features of the item to those in the current rule (which are based on items examined prior to the current item) and make an appropriate classification. Feedback then might result in the rule being modified. In particular, if an item were incorrectly classified, features might be added to the rule so that it will not apply to the incorrect category. Alternatively, features might be deleted from the rule so that it applies to the correct category. The process is repeated for the next training item.

According to this view, the knowledge-driven process and the standard learning model work relatively independently of each other. As each training item is presented, the standard model would “call” the knowledge-driven process, which would infer and/or select relevant features of the item. The learning algorithm of the standard model would then be applied without modification.

The findings of this experiment, however, highlight a basic limitation of these models. Namely, they assume that during learning, the knowledge-driven process is stable. However, our results suggest that this process can change during learning, in response to feedback. In other words, aspects of the learning algorithm may affect the knowledge-driven process: The modules are not independent.

In particular, some of our results suggest that prior knowledge interacts with feedback about classification to make features accessible to the learner. That is, people infer additional features about an item, using a combination of prior knowledge and feedback. One finding that suggests this interaction is feature reinterpretation. Consider the previous example in which a person thought a drawing was done by city child because it depicted a “city uniform,” but reinterpreted the clothing as a “farm uniform,” when told the drawing was done by a farm child. Clearly, the person uses prior knowledge to infer features about the drawing (the features are related to people’s expectations that farm and city children will draw people from rural and urban environments, respectively). At the same time, feedback affects this inference process. Initially, the person believed that the drawing depicted a “city uniform” but now believes that it depicts a “farm
uniform." Features are not only being added or deleted from rules: The processes for determining these features are being modified. A person is learning the features themselves. Presumably, given another similar drawing, the person now will be more likely to infer that the drawing depicts a "farm uniform" rather than a "city uniform."

A second related finding, which suggests that interaction, is "shifting the criteria for feature interpretation." Consider another previous example in which a subject believed that a drawing was done by a creative child because it was detailed. After receiving negative feedback, the person adjusted the criteria for interpreting a drawing as detailed. Given another similar drawing, the subject probably would not interpret it as being detailed. This strategy of shifting the criteria for feature interpretation again illustrates that prior knowledge coupled with feedback affects which features are inferred. In this example, the subject's prior knowledge makes accessible the notion of detail, and feedback is used to determine its applicability to a drawing. Note that feature reinterpretation also involves changing the criteria for feature interpretation. Presumably, when subjects change their interpretation of the data from one feature to a second feature, they also change the criteria they use to interpret both features.

Findings from this experiment also suggest that people use theories to construe the data as having one feature rather than another. In particular, people with different theories construed the same data as having different features (ones that were consistent with their theories). For example, a subject in the farm/city group interpreted the movement depicted in Drawing 5 of Figure 4 as "climbing in playground," whereas someone in the creative/noncreative group interpreted the same movement as "dancing." These findings are incompatible with the selection model because it assumes that theories function to pick out the important features (and therefore, that the features exist prior to the theories).

People's classification of some of the drawings also presents problems for the selection and addition models. Recall that for four of the drawings, people overwhelmingly classified them into one category and not into the other (see Table 5). For example, 95% of the subjects in the creative/noncreative groups classified Drawing 5 (shown in Figure 4) as "done by a creative child." Given the counterbalancing that was used in this experiment, however, the selection and addition models fail to predict that there will be a nearly unanimous consensus among people about an item's category membership. According to these models, an item is classified into the category whose corresponding rule applies to that item. These rules are determined by the standard model operating over training items. Therefore, any preferential classification of a drawing is based on learning.

To illustrate the failure of the selection or addition model to predict nearly unanimous consensus, consider in detail the classification of Drawing 5 by
the creative/noncreative group. As a result of counterbalancing, half of these subjects were given feedback that the drawings in Category 1 of Figure 4 were done by creative children. These subjects overwhelmingly classified Drawing 5 as "creative" and effectively placed it with the other items in Category 1. Therefore, according to the addition model, Drawing 5 should share more features with the items in Category 1 than with the items in Category 2, which accounted for why it was placed into Category 1 (and classified as creative). Given this reasoning, it also follows that the other half of the subjects should also have placed Drawing 5 in Category 1. (These subjects were given feedback that the drawings in Category 2 of Figure 4 were done by creative children.) However, these subjects also overwhelmingly classified this drawing as "creative," and effectively placed it with the items in Category 2.

Apparently, subjects strongly believe that some features are explicitly associated with a category, prior to having seen any of its members. This finding suggests another important way that knowledge effects differ from those suggested by the selection or addition model. Knowledge generates expectations about which features are linked to specific categories. So, for example, subjects not only think that "detail" is relevant to distinguishing creative drawings from noncreative ones but they also believe that it is the creative drawings that will show detail. Therefore, it is not sufficient for knowledge to add, select, or weight features because of their general relevance to category learning.

Incorporating category-specific beliefs into a selection or addition model would only require minimum modifications. As suggested later, however, such a step is more in the spirit of a learning model in which knowledge and experience strongly interact and are tightly coupled. Neither the addition or selection models capture how knowledge affects learning. To account for knowledge effects, one must abandon the notion that knowledge and learning components are modular and operate (relatively) independently of each other.

**GENERAL DISCUSSION**

These studies suggest a view of learning with prior knowledge that is fundamentally different from the standard view of category learning. Consequently, our findings cannot be accounted for by straightforward extensions of this view that would preserve the standard model but allow for knowledge only to select or add features to training items. These studies also go beyond previous work that has examined the role of prior knowledge in category learning. This work has emphasized the facilitatory or inhibitory effects of knowledge on learning. A number of researchers have implicitly assumed that the learning process involved does not differ from that of the standard model (e.g., Lien & Cheng, 1989; G. Murphy, 1993; G. Murphy & Wisniewski, 1989; Pazzani, 1991; but, see Wattenmaker et al., 1986, for an exception).
Given poor theories or expectations about a category, learning involves an interactive specialization process. In particular, people's intuitive theories and expectations may activate abstract hypotheses about the category. Learning then involves specializing these features, that is, determining other features that support or provide evidence for these more abstract features. Importantly, our studies suggest that theory- or knowledge-driven processes interact and are tightly coupled with data-driven processes in determining how and which abstract features are specialized. Knowledge-driven processes influence data-driven processes and vice versa. These processes are not modular.

This type of learning contrasts with the standard view in cognitive psychology. According to the standard view, learning only involves finding some set of features that tend to be shared by the members of a category and that distinguish the members from other categories. In one sense, this view is compatible with our proposal in that the abstract features that people specialize may be shared by members of a category and may distinguish the category from other categories. However, our proposal suggests that learning involves much more than finding features common to a category that distinguish it from other categories.

The results have a number of implications for how one might integrate prior theories and expectations with experience and data during learning. We will look at this issue from three perspectives. First, we revisit the standard model and evaluate its basic assumptions in terms of our findings. Next, we describe some aspects of a psychological model in which prior theories and experience are tightly coupled during learning. Finally, we consider AI systems that have integrated prior knowledge and experience in learning. Currently, they are the only concept-learning systems in which prior knowledge plays a role. We will show that these models are either variants of the addition model or that the interactions between prior knowledge and experience are not tightly coupled.

**The Standard Model Revisited**

At the minimum, our results suggest that straightforward ways of incorporating prior knowledge into the standard model will not account for how people learn about categories. Even more fundamentally, however, we believe that a number of our findings go against the basic assumptions of the standard model (listed in Table 1). In the following, we describe these basic assumptions and examine them in light of our results and some intuitions.

**Learning Involves Selecting Features from a Space of Prespecified Unambiguous Features**
The standard model presupposes the existence of the features over which learning operates. Category items are described by a space of clearly demarcated features. Learning then involves selecting a subset of features from this
space that are relevant to the category. The assumption of a well-defined feature space is also carried through in psychological studies designed to evaluate such models. Typically, the experimenter explicitly provides the subject with the features that describe the training items (e.g., by providing linguistic descriptions of the features). Alternatively, the experiments involve items consisting of a few basic, perceptual dimensions (e.g., color, shape, size) in which there is implicit agreement between the experimenter and subject on the features that characterize these items.

In some contexts, the assumption that features are a given may be reasonable. For example, experts in a domain may be able to provide a learning system with training items and their associated features. In fact, experts provided AQ11 (Michalski & Chilausky, 1980) with descriptions of soybean plants and their diseases. Given these descriptions, the system learned rules for diagnosing soybean diseases that outperformed those of the experts.

In most contexts, however, it is generally true that features are not "just out there," ready for a learning system to operate over them. Features undoubtedly are determined by the theories and expectations about the data that people bring to a learning situation. This point becomes strikingly clear when one considers some of the interpretations of the children's drawings in the Experiment 2. In particular, people with different theories interpreted the same part of a drawing differently and in a way that was consistent with their theories. Our study is not the first to demonstrate such theory-influenced coding of the data: Many examples abound in nonlearning contexts such as the social domains (e.g., Nisbett & Ross, 1980) and science (e.g., Brewer & Lambert, 1991). For example, early drawings of Saturn depicted it with satellites rather than rings, most likely because of the belief that planets had satellites. When astronomers developed new theories about Saturn, later drawings depicted the rings (Brewer & Lambert, 1991).

In the standard model, the processes that produce the features are not addressed. In defense of this practice, one could argue that these processes should be carried out by a separate module. That is, one might assume that features were given to the learning component by a perceptual system interacting with knowledge structure (as implied in the previous examples). The details of this process would not be relevant to learning. Other researchers (like those studying vision) would flesh them out. Eventually, a learning system operating in the real world would interface with such a process.

On the other hand, findings from Experiment 2 suggest that part of learning involves learning these features. Specifically, in the course of modifying their classification rules, people sometimes reinterpreted the features of a drawing and changed their criteria for deciding that a drawing had a particular feature. In other words, part of learning about a category involved hypothesizing the features that described an item and modifying those hypotheses about features. In contrast, the standard model assumes that the
features are given and only hypothesizes which features are relevant to the category. With the exception of some AI systems (to be described), researchers have generally ignored this aspect of learning.

Another claim is that the process of figuring out the features results from a tightly coupled interaction of prior theories and experience. When people determine the features that describe a category item, they may rely not only on what they “see” in the item, but also on what they have seen in prior category items, and on general knowledge not specific to the category.

**Learning Involves Selecting Features Based on Statistical Regularities**

Most theories of category learning have emphasized statistical dimensions of features. Typically, the significance of a feature is based on its frequency among the members and nonmembers of that category. It seems clear that a model of category learning should be sensitive to feature frequency. Feature frequency is important because it is strongly related to two basic functions of concepts: classification and prediction. One can use features that occur among the members of a category, but that infrequently occur among non-members to classify a novel item into the category. For example, the feature “has feathers” frequently occurs among birds and infrequently occurs among things that are not birds. Therefore, given that some item has feathers, one might reasonably assume that it is a bird. Also, given knowledge about an item’s category membership, one can predict features about the item that frequently occur among category members. So, given that some item is a bird, one might reasonably predict that it “lays eggs.”

On the other hand, category learning involves much more than attending to the frequency of features among the members and nonmembers of the category (see, also, Keil, 1989; Medin & Shoben, 1988; G. Murphy & Medin, 1985; Rips, 1989). As noted, our experiments suggest that learning involves the instantiation or specialization of abstract features. In this case, people’s prior knowledge activates a space of abstract features and hypotheses. People than are biased to search for evidence among the category items that supports these features or clarifies more specifically how the abstract features are instantiated in the items. A strong intuition is that much of adult learning involves the instantiation of abstract expectations about a category. For example, it frequently seems that people are faced with the task of learning how to operate human-made objects (televisions, computers, musical instruments, etc.). Learning involves having an expectation about the intended function of an object and figuring out how to carry out (i.e., instantiate) that function.

One effect of instantiation is that it sometimes allows people to treat features as equivalent, or similar when in other contexts they are usually considered different. For example, in a neutral context, the features, pockets and collars appear quite different. However, given the theory-activated
expectation that drawings will show *detail*, these features now appear to be similar because they instantiate this notion of *detail*. As another example, in a neutral context, an object that "emits microwaves" might appear quite different from one that "contains poison." But, given the expectation that these objects are "used for killing bugs," the features may now appear to be similar because they are examples of methods for killing bugs.

Importantly, models that treat features as equivalent in this manner may be capable of learning some categories that the standard model cannot learn. For example, consider the category of "drawings done by emotionally disturbed children" and suppose that its members are distinguished from other categories of children's drawings by the presence of abnormalities. It is our intuition that the space of possible abnormalities manifested in a child's drawing is *infinite*. Such a space might include "missing foot," "missing nose," "disproportionate legs," "an arm poorly connected to the body," "a tiny drawing," "a very large drawing," "a shaded figure" to name just a few (these features are actually diagnostic of children with emotional disorders; see Koppitz, 1984). Unless a model could treat these different features as examples of the abstract feature "abnormality," it could not learn this category.

### Classification Is Based on a Matching Function Involving Syntactic Identity

In the standard model, items are classified by comparing the item's features with those of the category representation. The certainty or degree that an item belongs to a category is some function of the item's features that match and mismatch those in the category representation. For example, in prototype models, an item's features are compared with those of the prototype (which typically includes the most frequent features of the category members). In a number of these models, the certainty that the item belongs to the category is an additive function of the item's features that match and mismatch those of the prototype.

In calculating matches and mismatches, one issue is how to align the features of the item with those of the category representation in order to compute matches and mismatches. In other words, given a feature in the item, which feature in the representation does it match or mismatch? Typically, most models assume that category items and representations of categories consist of attribute-value pairs (e.g., color: red, height: 6 ft). Given this assumption, an attribute-value pair of the item matches a *syntactically identical* attribute value in the representation (e.g., color: red in the item matches color: red in the representation). An attribute-value pair of the item mismatches an attribute value in the representation if they share the same attribute but a different value on that attribute (e.g., color: red in the item mismatches color: purple in the representation).
On the other hand, it may be that classification sometimes involves more than this straightforward view of alignment and computing of syntactic matches and mismatches (see also G. Murphy & Medin, 1985). For one thing, in other cognitive processes such as interpreting analogies and making similarity judgments, feature alignment is more complex (Gentner, 1989; Markman & Gentner, 1993a, 1993b; Medin, Goldstone, & Gentner, 1993). More specifically, in classifying an item, people may make inferences about the item and be selective about which features are aligned in the item and category representation. Although we have no direct evidence from these studies that supports this view, we have a strong intuition that this is the case. Several examples illustrate this intuition. First, consider some drawings of people done by creative children that can be characterized as “detailed.” Suppose that a new, detailed drawing is presented with the feature “has earrings” and that none of the previous drawings done by creative children contain earrings. Clearly, the category representation for drawings done by creative children will not contain the feature “has earrings” and is unlikely to contain the feature “does not have earrings.” Therefore, in the standard model, it is unclear whether “has earrings” matches or mismatches a feature in the category representation. Our intuition suggests that people may use “has earrings” (along with other features) to infer that the drawing is “detailed,” and thus classify the drawing as creative. Here, people do not consider “has earrings” as matching or mismatching a feature in the category representation. Rather, they infer that the feature supports a more abstract feature of the item, which is aligned and matched with its counterpart in the category representation.

As a more extreme example, again consider some drawings of people done by creative children, which can be characterized as “detailed.” Suppose that a new drawing of a tree is presented with finely textured bark and intricate leaves. We suggest that people will once again infer the drawing is “detailed” and thus classify the drawing as creative. Here again, people do not consider the features “finely textured bark” and “intricate leaves” as features that match or mismatch ones in the category representation. Rather, they infer that these features support a more abstract feature of the item, which is then aligned with its counterpart in the category representation.

Tightly Coupling Knowledge and Experience During Learning
Our findings imply that knowledge and experience closely interact during concept learning. In particular, prior knowledge influences the interpretation of experience, and experience influences prior knowledge. In this section, we propose a model of learning in which the interaction of knowledge and experience is tightly coupled. A key assumption of the model is that not all features are simply a given and that an important part of learning is determining the features. Furthermore, people rely on a variety of sources
of information to determine features (see also Wisniewski & Medin, 1994). The model is speculative and more research is needed in order to elucidate the precise effects of knowledge on experience and vice versa.

To illustrate the model, we will focus on a particular learning context: namely, the situation presented to subjects in Experiment 2. In that study, items belonging to two categories were presented sequentially. A subject’s task was to acquire a concept that distinguishes one category from the other. On a given trial, learning can be viewed as a function of a variety of sources of information. As Figure 5 shows, these sources of information include prior knowledge or expectations about the category, general knowledge, and three sources of empirical information: previous items, feedback about a subject’s classification decisions, and the item currently presented.

To illustrate the tightly coupled interaction of prior knowledge and experience, consider how a person might determine a feature of the current item, given different positions in a sequence of training items. Suppose that the current item is Drawing 3 in Figure 4 and that it is the first item presented in the sequence. Low-level data in this item such as the vertical orientation of the dots and their location in the center of the upper torso is consistent with both the feature “tie” and “buttons” (as suggested by subjects’ rules given in Experiment 2.) In other words, the data are ambiguous with respect to indicating these features. People with different prior knowledge about the category will interpret these data differently. For example, prior knowledge that the drawings have been done by creative and noncreative children may activate an abstract feature such as “detailed.”
As a result, a person may be likely to interpret the drawing as having "buttons" rather than a "tie" because "buttons" provides more evidence for "detail."

Now suppose that the same item was preceded by other items instead of appearing as the first item in the sequence. The likelihood that a person will interpret the drawing as having "buttons" may further change, depending on the frequency of "buttons" in the previous items and the extent to which "buttons" helps predict category membership. For example, if "buttons" has provided evidence for "detail" in many previous creative drawings, then a person should be more likely to interpret the item as having buttons than if it had been the first item in the sequence. This likelihood may change for at least two reasons. First, a person is more (less) likely to search for "buttons" to the extent that the feature has predicted (not predicted) category membership (as most models of category learning would predict). On the other hand, a person may also assign more (less) weight to the data as evidence for "buttons" if similar data have been consistent (inconsistent) with the presence of "buttons" (as current models do not predict). In this case, both a person's prior expectations and the presence of a feature in previous items influences the interpretation of a feature in the current items. (For related examples in nonlearning tasks, see Asch, 1946; Bugelski & Alampay, 1961; G. Fisher, 1968; Medin et al., 1993.)

There are more complex ways that knowledge and experience interact in interpreting features. First, the interpretation of a feature may also be influenced by other features in the item. For example, suppose that a number of detailed, creative drawings contain "buttons," "shoelaces," and "eyebrows." Again suppose that the current item is Drawing 3. A person may be more likely to interpret the drawing as having "buttons," if it also has "shoelaces" and "eyebrows." (Lesgold et al., 1988, provided evidence for this phenomenon in the domain of x-ray interpretation.) Second, there is also evidence in the domain of dermatology that the interpretation of a feature is partially influenced by categorization of the item (Norman, Brooks, Coblentz, & Babcock, 1994).

These experiments have examined situations in which prior knowledge has been activated by meaningful category labels. It also seems that the data themselves could activate prior expectations. For example, a drawing of a person holding a pitchfork could remind a person of farms and bias interpretation in other aspects of that same drawing (e.g., the clothing might be construed as less fashionable).

In general, it may be best to view the presence of a feature as being determined by some critical amount of evidence. The evidence for a feature is a function of the information provided by prior expectations, previous items, and the current item (including other features in the item). The contribution of these different sources of knowledge may vary and sometimes one may
dominate others. For example, in the current context, suppose you were told that you would see drawings of elephants and then were presented with a number of such drawings. Now, you are presented with a drawing of a person. It is highly unlikely that you would interpret the person’s nose as an elephant trunk.

Importantly, learning sometimes involves determining the nature of the evidence for a feature. For example, in learning that creative children draw detailed pictures, one has to learn what constitutes evidence for the feature “detailed.” A “shoe” may be evidence for “detailed” in a drawing done by a creative 4-year-old but not in one done by a creative 12-year-old. In contrast to this view, most learning systems assume that the features are unambiguous and present with certainty.

**AI Systems that Integrate Knowledge with Experience**

In contrast to concept-learning models in cognitive psychology, a number of machine-learning systems integrate prior knowledge with experience (e.g., Flann & Dietterich, 1989; Lebowitz, 1986; Mooney & Ourston, 1989; Shavlik & Towell, 1989; Towell, Shavlik, & Noordewier, 1990; Yoo & D. Fisher, 1991). These systems typically represent prior knowledge in the form of inference rules. A number of systems are variants of the addition model and add features to items via deductive rules of inference. Some systems assume that their rules are only approximately correct and that learning involves rule modification (and therefore, determining the nature of the features). In this section, we describe these systems and evaluate their potential as process models of human learning. These systems are impressive examples of integrating prior knowledge with experience. We will suggest, however, that they do not tightly couple the interaction of knowledge and experience in the manner that we have described.

In general, AI systems that integrate knowledge-based and empirical learning use some variation of the standard model of learning as the empirical component and use a knowledge-based component that is popularly known as *explanation-based learning* or EBL (for details, see Dejong, 1988; Dejong & Mooney, 1986; Ellman, 1989; Flann & Dietterich, 1989; Mitchell et al., 1986). Figure 6 illustrates the basic components and operation of an EBL system. Such a system has a preexisting body of knowledge called a **domain theory**: a set of well-defined, deductive inference rules (e.g., “If item has a handle, then item is liftable”). Input to an EBL system typically consists of a **training item**, described as a set of well-specified features (“lightweight,” “white color,” “handle,” etc.) and a **target concept** (e.g., “item is drinkable-from”).

Typically, an EBL system operates over a set of training items. For each item, the system uses its domain rules to prove or explain deductively why a particular item is an example of the target concept. The domain theory rules
domain theory rules

Rule 1: IF Item is liftable AND Item is open AND Item is stable AND Item is a container THEN Item is drinkable-from

Rule 2: IF Item is light AND Item has a handle THEN Item is liftable

Rule 3: IF Item is concave THEN Item is open

Rule 4: IF Item is concave AND Item has an upward orientation THEN Item is a container

Rule 5: IF Item has a flat bottom THEN Item is stable

Figure 6. Illustration of explanation-based learning (EBL).

indicate which features must be present in an item in order to infer additional features of the item. Adding these features involves the matching of features in an item to those in a rule and the inferring of additional features. The process is recursively applied until the system constructs an explanation that links features in the training items to intermediate features, which in turn are linked to the target concept. As shown in Figure 6, the explanation is a tree structure whose root is the target concept and whose leaves are a subset of the features of the training item. The intermediate nodes of the tree are features that the explanation-based component has deduced about the item or added to the item by using its domain theory. Once an explanation is constructed, it is generalized so that it potentially applies to items beyond those in the training set (Mooney & Bennett, 1986, described one common generalization algorithm).

Traditionally, EBL has been viewed as a method of knowledge compilation. For example, upon constructing and generalizing an explanation, an EBL system could extract a rule whose preconditions are the leaves of the
generalized explanation tree and whose conclusion is the root (the target concept). So, in the current example, the system could generalize and compile the explanation into the rule "if an item is 'lightweight,' 'has a handle,' 'has a flat bottom,' 'is concave,' and 'has an upward orientation,' then it is 'drinkable-from.'" Given a new item with these features, the system could infer that it was "drinkable-from," by using the rule, rather than by reconstructing the explanation. If an EBL system also performed some task (such as problem solving or object recognition), then this type of compiled knowledge may sometimes improve performance (as in SOAR; Laird, Rosenbloom, & Newell, 1986). However, the effects of compiled knowledge on performance are not straightforward and some studies have shown that adding compiled knowledge to a system can degrade its performance (e.g., Minton, 1988). More recently, EBL has been viewed as both theory-guided specialization and as induction rather than solely as a form of knowledge restructuring (Flann & Dietterich, 1989).

Addition Systems

Addition systems fall into two broad categories. Some systems learn by processing training items with an empirical component that then sends its output to a knowledge-based component. More commonly, there are systems that use these components in the opposite order: They process training items with a knowledge-based component that then sends its output to an empirical component (see Mooney, Ourston, & Wu, 1989, for a review).

Lebowitz (1986) detailed a system that uses an empirical component followed by a knowledge-based component. The system has been studied in the domain of congressional voting. Using a conceptual clustering system called UNIMEM, the system detects features that frequently occur across training items (e.g., all the members who voted for a particular bill are Southern democrats). It then attempts to explain these commonalities with its domain theory (e.g., why did Southern democrats all vote for the particular bill). The explanation effectively determines those common features of the training items that are causally relevant to the outcome, and those that are purely coincidental. Lebowitz suggested that this approach apparently improves the efficiency of the explanation process.

Several models that operate in an opposite manner include "Induction Over the Unexplained" or IOU (Mooney & Ourston, 1989), "Induction over the Explained," or IOE (Flann & Dietterich, 1989), and EXOR (Yoo & D. Fisher, 1991). IOU first applies EBL on training items. Features that do not enter into the explanations become input to an empirical component, which is used to detect those features that tend to occur across the items. The target concept is then augmented with these unexplained features. One reason this approach is important is that categories may have predictive features that the domain theory does not select as relevant. IOE constructs
explanation trees for a number of training items and employs empirical learning to detect frequently occurring substructures within the trees. It also uses empirical learning to detect certain features or patterns of features that occur in the explanations. These "common coincidences" in explanations are then incorporated into the target concept. For example, the feature "made of plastic" might be incorporated into the target concept "drinkable from" if this feature was a common component of a series of explanations involving particular cups. Like IOE, EXOR constructs explanation trees for a number of training items and employs empirical learning (i.e., COBWEB; D. Fisher, 1987) to detect frequently occurring substructures within the trees. EXOR uses these frequently occurring substructures to organize the explanations hierarchically so that they can be efficiently accessed in future situations.

In these addition systems, the interaction between knowledge and experience is indirect and operates in one direction. Typically, the first component acts as a filter for the input to the second component by reducing the number of features that the second component processes. For example, in IOE and EOR, the explanation-based component first processes the items and then passes the explanation trees to the empirical component. In Lebowitz's (1986) system, the empirical component processes the items and then passes a subset of their features (i.e., the frequently occurring features) to the explanation-based component. In these examples, one component influences the other but not vice versa. The interaction of knowledge and experience is not tightly coupled.

One consequence of this indirect interaction between prior knowledge and experience is that the application of inference rules is not influenced by their use in constructing explanations for previous training items (but, see Tecuci, 1991, as one exception). Suppose, for example, that a domain theory rule had been involved in a number of explanations of previous training items. An EBL system would not be biased to favor or select this rule in forming an explanation of the current training item. In contrast, it seems clear that people's preference for one rule versus another depends on its successful application on previous trials. That is, people empirically monitor the success and failure of their prior knowledge in explaining the training data.

Rule-Modification Models
The addition systems described previously tend to operate in domains that are generally well defined and characterized by correct rules of inference (e.g., IOE has been applied to the chess domain). As a result, the rules remain unchanged as explanations are constructed over a training set. For example, a rule like "If item has a handle, then item is liftable" (shown in Figure 6) would be unmodified in the course of forming explanations for a
series of items. At least several systems, however, explicitly modify their rules in the course of learning.

Mooney and Ourston (1992) have developed a system called "Explanation-based and Inductive Theory Extension and Revision" (EITHER), which refines a domain theory that is only approximately correct. The system first applies EBL to a series of training items. Because its domain theory is only approximately correct, EITHER may fail to explain why some items are examples of the target concept and incorrectly explain why some items are examples of the target concept (when they are not). An empirical component (ID3; Quinlan, 1986) then focuses on these explanation failures to detect errors in the domain theory. EITHER modifies rules by adding antecedents to them (making the applicability of a rule more specific) or by removing antecedents (making the applicability of a rule more general). It can also acquire new rules.

Note that in modifying rules, EITHER is learning features. Therefore, it goes beyond the previous models that use preexisting, unmodifiable deductive rules to add features. On the other hand, the interaction of prior knowledge and experience is not tightly coupled in this system. Like the addition models, EITHER first applies its prior knowledge to all training items and then uses empirical learning. Also, the application of this prior knowledge is not influenced by its prior use. That is, the system's application of a domain rule to the current training item is uninfluenced by its previous success or failure in constructing explanations for previous items.

Another rule-modification system called "Knowledge-Based Artificial Neural Networks" (KBANN), uses its approximately correct domain theory to construct a connectionist network, whose nodes, connections, and weights are isomorphic to the domain theory rules (Shavlik & Towell, 1989; Towell et al., 1990). In processing a series of training items, the system refines its domain theory. In a series of simulations, KBANN was more accurate at classifying certain sequences of DNA nucleotides than a standard empirical model (ID3) and a connectionist network whose structure was not determined by a prior domain theory.

Like many connectionist networks, the system's weights are modified after the presentation of a training item. As a result, there will be a change in the probability that a given node will become activated when the next training item is presented. Hence, the system's classification of the current item is influenced by previous items (unlike EITHER and the addition models). However, KBANN does not tightly couple the interaction of prior knowledge and experience. The system uses prior knowledge to construct the topology of its connectionist network. However, the network does not have connections from high-level nodes (corresponding to the prior knowledge) to lower level nodes that influence the activation of the lower level nodes during learning (as in, for example, interactive word-perception model; McClelland & Rumelhart, 1981).
The Need for General Domain Theories in EBL

Our intuition is that people's prior knowledge about some domains may be rather general and that the specific aspects of the domain cannot always be anticipated. As we previously suggested, a domain such as "drawings done by emotionally disturbed children," may be characterized by an infinite number of possible features that, in principle, could be evidence for emotional disorders. An intelligent system cannot possibly represent all of these features prior to having interacted in the domain. As a result, it is unlikely that learning could always "get off the ground" by the application of rules that straightforwardly permit inference of a feature from other features (as in EBL). A challenge to any system will be how to construct more general mechanisms that can recognize unanticipated features and hypothesize a link between them and other features (e.g., hypothesize that "pictures of slugs drawn on a shirt" is evidence for "unusual" in an emotionally disturbed drawing).

Furthermore, it is impossible to anticipate all of the possible domains for which one's prior knowledge will be appropriate. For example, one domain investigated in these experiments was "drawings of people done by creative children." However, the knowledge that people used in learning about this domain may have also been relevant to domains such as "drawings of people done by creative adults," "drawings of trees done by creative children," and so on. It is highly unlikely that people have rules specific to these domains. Another challenge to any system will be how to construct specific rules from general knowledge for a domain that it could not anticipate.

Summary

Our studies have shown that a simple manipulation (varying the label associated with a category) dramatically influences rule learning. The effects of this influence cannot be captured by treating empirical learning models as separate modules that are only indirectly influenced by prior knowledge. At least in some domains, learning is not a process in which prior knowledge first exerts its influence on data and then empirical learning takes over afterward. Rather, it is a process in which prior knowledge and experience are closely interwoven. Prior knowledge may construe the experiences over which learning takes place. On the other hand, experiences may affect which prior knowledge construes experiences and how strongly it construes those experiences.

This research has also demonstrated that different sources of information may interact in various ways to determine the features over which learning takes place. This finding undermines a basic assumption of many concept-learning models that learning operates over a space of well-defined features.

Many issues still need to be explored. It is unclear exactly how the various sources of knowledge are combined and weighted in determining
features. Future work needs to carefully manipulate these sources of knowledge in order to derive more quantitative predictions about learning. In any event, we hope that this article illustrates the importance of viewing learning as theory instantiation and as a highly interactive process involving the mutual influences of prior knowledge and experience.

REFERENCES


