Observational Learning From Internal Feedback: 
A Simulation of an Adaptive Learning Method

DORRIT BILLMAN
University of Pennsylvania

EVAN HEIT
Stanford University

Much natural learning occurs by observation without explicit feedback or tutoring, yet few models of learning address this class of tasks. Further, many natural cases of observational learning are complex, and efficient learning seems to demand strategic learning procedures. The present work adopts a design perspective and asks what learning mechanisms would be both useful and feasible for natural induction. Work on closely related learning problems is briefly reviewed and a model for observational rule learning is proposed and simulated. The model extends learning mechanisms developed for explicit learning with feedback to less structured, observational tasks. In particular, the focused sampling mechanism, which is an extension of the attentional learning procedure developed by Zeaman and House (1963), is introduced. The operation of an attentional learning procedure is less clear when extended to learning without feedback, so a simulation was done to evaluate the performance of the model. A series of simulated experiments were run, comparing performance of the learning model with and without the focused sampling component. We evaluated whether and when focused sampling benefits observational learning, investigated the effects of different distributions of systematic and unsystematic features, and compared observational learning to learning with feedback. Results of the simulation suggest that focused sampling does benefit learning, that benefit increases with the complexity of the learning task, and that learning with and learning without feedback exhibit differences in how each is affected by changes in the learning problem. Suggestions about the relation to human data are offered.

INTRODUCTION

Many of the categories and rules of language, of social interaction, and of behavior of everyday objects are learned in untutored, observational con-
tions. In the absence of explicit feedback, the successful learner must be sensitive to the structure available from observation of the input examples themselves and be able to extract and represent relevant structure. Indeed, much of our structured knowledge stems from observational learning. By structured knowledge we mean knowledge which includes abstractions such as rules and categories. The research reported here focuses on rule learning, but we assume that category and rule learning are intimately linked and that psychologically natural categories, such as NOUN, TEACHER, and DOG, coordinate a set of rules or regularities about an interrelated set of properties. By observational learning we mean learning without explicit external feedback and without tutoring about category membership or adherence to a rule.

The bulk of research on category and rule learning has focused on learning from feedback. "Feedback," unmodified, and "external feedback" both refer to explicit, external information that 1) provides success criteria for a particular learning problem and 2) is taken as such by the learner. Tasks differ in the availability of feedback. In the most commonly investigated type of task, the learner is provided with instances which belong in a target category, instances which do not belong, and discriminative feedback information identifying which are which (Anderson, Kline, & Beasely, 1979; Bruner, Goodnow, & Austin, 1956; Medin & Schaffer, 1978; Reed, 1972).

In the second type of task, instances from multiple categories are provided but there is no feedback assigning instances to categories. Learners might (Fried & Holyoak, 1984, Homa & Cultice, 1984) or might not (Brooks, 1978) know there are categories to be learned, but they do not know the assignment of instances.

In the third type, the learner is initially provided with instances of only one category. Hence, there can be no discriminative feedback distinguishing among contrasting categories which might help identify the basis for category membership. In tasks where discriminative feedback is available, it provides a success criterion or goal and can be used by a control process to guide learning. From the learner’s perspective, provision of explicit, external feedback narrows the learning task by prioritizing one aspect of input which is to be predicted, such as whether an instance is a category member or not. In the third type of task which lacks discriminative feedback and contrasting categories, part of the learner’s problem is to figure out which aspects of the input are useful to predict. We will address this third type of learning here.

Our research investigates learning mechanisms which would contribute to successful observational learning. Specifically, we simulate mechanisms for using match and mismatch of expectation to generate internal feedback. As psychologists we are also interested in whether people do use the simu-
lated mechanisms in conscious rule learning and hypothesis testing, in more passive and implicit learning, or in both. The present study does not answer this question, but we return to it in the discussion.

This paper presents a model of learning from observation of the correlational structure afforded by input and reports results of a simulation. Two major issues must be addressed by any model of this type. First, how are correlational rules learned at all in the absence of external feedback? Second, how can such learning be made tractable for large inductive problems? The present model combines and extends components from established models to address these issues.

Learning Covariation Rules without External Feedback

Some psychological models of rule and concept learning address learning without feedback, and some address learning correlational rules, but few established models, in psychology or artificial intelligence (AI), address both. Models which do not depend on feedback and which specify a learning mechanism, learn by similarity clustering and do not acquire rules about covarying features or correlational rules (Fried & Holyoak, 1984; Gibson, 1968; Michalski & Stepp, 1983). Models which learn correlational rules use explicit, discriminative feedback as critical components of their learning mechanisms (Anderson, Kline, & Beasely, 1979; Bruner, Goodnow, & Austin, 1956; Hayes-Roth & Hayes-Roth, 1977).

Two models, one in language acquisition (Maratsos & Chalkley, 1980) and one rule learning system from AI (Davis, 1985), do combine learning about feature covariation with independence from external feedback, and provide important precedent for the current approach. Maratsos and Chalkley propose that learning co-occurrence properties without external feedback may form the basis for learning syntactic categories. They use learning noun gender categories as an illustration and develop an account of how gender classes might be learned by coordinating rules about co-occurrences of gender-related markings.

Davis (1985) has developed a learning program which discovers, not categories, but conditional rules among a large set of potentially relevant features. The system conjectures that the presence of one feature, for example, being a prime number, is likely given the presence of another, such as being odd. The system can operate in "self-test" mode, generating and testing such conjectures. If a predicted property is found given the predicting feature, the predictive rule's strength, or estimated conditional probability between the two features, is increased.

As in Davis's program, our learning system samples information about one set of features, projects the expected values of a different set of features, and revises rule strength in light of the match or mismatch to observed values. Our model differs from Davis's in its procedure for making the in-
uctive learning more tractable. If complex systems of rules are to be learned in a practical amount of time, generation and testing of conditional rules must be guided by some strategic component. Focused sampling is one such strategic component, or adaptive plan.

While many connectionist models depend heavily on predicting a prespecified feedback feature, some are designed as auto-associators or regularity detectors (McClelland & Rumelhart, 1986; Rumelhart & Zipser, 1986), and these can learn correlational patterns without explicit outside feedback. Unlike our work, connectionist research has not emphasized strategic allocation of attention to make complex learning tasks easier. Nor has it emphasized the role of feedback in changing the definition of the learning task. Connectionist models only adjust weights on connections between features, while focused sampling adjusts feature weight and relational weights separately.

Making Induction Tractable
Successful guidance to the learning procedure must come both from restricting what can be represented in the language and from learning procedures which use partial knowledge to direct ongoing learning. Work on complex learning of syntax (Chomsky, 1965; Pinker, 1984; Wexler & Culicover, 1980), of categories (Keil, 1979), and of subtraction (Van Lehn, 1983) has focused on restricting the representation language, rather than on guiding the application of inductive learning procedures. However, for many problems an intractably large space of possibilities still remains.

Holland's research (1975) introduced and analyzed adaptive plans which operate on the basic processes of rule generation and testing to use available knowledge to guide the course of learning. The work of researchers from several disciplines is relevant to understanding the use of adaptive plans. In philosophy, Goodman (1983) has outlined how partial knowledge about higher-order or superordinate terms might guide further learning and induction. These ideas have begun to be applied to research on induction in psychology (Nisbett, Krantz, Jepson, & Kunda, 1983; Shipley, 1988a) and AI (Russell, 1986). In a similar vein, Grimshaw (1981) has suggested that the learning processes in language acquisition, not just innate universals specifying allowable representation, may contribute toward constraining language structure.

In the present model, focused sampling is a form of attentional learning which functions as an adaptive plan. Focused sampling guides the generate-and-test procedures to explore rules which are interrelated. The issue of attention allocation and attentional learning has been addressed in prior research on concept and rule learning. Trabasso and Bower (1968) suggested that learners in a simple, well-defined concept learning task showed evidence of attentional limits but not of learning how to allocate limited attention; rather, subjects sampled features independently from trial to trial. In contrast, Zeaman and other researchers provided models and evidence for at-
tentional learning by people (Fisher & Zeaman, 1973; Zeaman & House, 1963) and animals (Lovejoy, 1966). The established models of attentional learning address explicit learning of rules or categories where feedback is explicitly provided. Focused sampling begins with the Zeaman & House learning procedure and shows how this can be freed from dependence on external feedback; doing this requires a basic modification in the control of attentional learning.

The Zeaman models (Fisher & Zeaman, 1973; Zeaman & House, 1963) increase the salience of a stimulus feature whenever that feature is selected as a predictor and its prediction of the feedback is correct. Suppose the learner selects "color" as the predictor, observes the current instance is blue, applies a rule stating that blue objects are (or are not) category members, and makes the prediction that the object is in the category. If the category is as predicted, the salience of "color" is increased. This increase indirectly robs salience from other features, because salience is normalized across features. Thus, if one feature covaries with another, as would be the case if blue color and square shape were both cues to category membership, attending and learning about one predictive feature competes with knowledge of others. A competitive relation between correlated predictor features is also found in animal conditioning and results in blocking and overshadowing; learning about one feature is more difficult when that feature covaries with another predictor feature (Mackintosh, 1975; Rescorla, 1972). Learning about color interferes with learning about shape, if both covary with each other and predict feedback.¹ In the models for learning with feedback, predictor features which covary with each other are in competition.

A Model For Learning Rules Without Feedback

*Rule Generation and Testing with Internal Feedback.* We assume knowledge is represented as conditional rules, consisting of a given and a predicted set of feature values. A feature is a type of characteristic or dimension which can assume different values. This representational form of conditional rules is similar to that used in other models of concept learning (Anderson, Kline, & Beasely, 1979; Hayes-Roth & Hayes-Roth, 1977). In these models the predicted feature is always the category label, but it need not be in our model.

¹ In the Mackintosh model (1975), learning about one feature competes even more directly with learning about another covarying feature: change in feature salience only occurs when that feature is the best predictor of the feedback. Hence, if Feature A (color) is rated a better predictor, no attentional learning about a covarying Feature B (form) will occur. Attentional competition among features does not result from renormalizing the total attention to all features following an increase in attention to one feature but from precluding the eligibility of a feature to be used in attentional learning.
Our primary suggestion about the rule learning process is that predictive success of conditional rules can be used to generate internal feedback to guide learning. A simple version of the learning model uses rules initially formed by selective encoding of input and applies the following learning cycle. The learner (1) selects features for the condition part of a rule, (2) determines the values of these features in input, (3) makes a prediction about some other aspect of the input by selecting or forming a conditional rule using these feature values in the given part, and (4) tests the rule by comparing the predicted with actually occurring feature values. The match or mismatch between expected and observed values yields internally-generated feedback which can be used to guide rule learning. The internal feedback is used to strengthen or weaken a rule and to select which rules are most likely to be additionally tested. (A more complete version of the model would include selective generation and modification of rules and would use internal feedback to guide these mechanisms.)

Learning procedures which generate and modify rules or associations based on predictive success are widespread (Anderson, Kline, & Beasely, 1979; Rescorla, 1972; Rumelhart, Hinton & Williams, 1986). The internal feedback model suggests that control from predictive success need not derive only from explicit feedback, but can be based on predicting dependencies among any features of input.

**Focused Sampling.** Focused sampling is an attentional learning procedure which can be added to the basic mechanisms for generating and testing rules to guide their application. Focused sampling allocates limited learning resources to a subset of rules, those deemed more likely to be valid. As such, it forms a simple adaptive plan. It uses internal feedback to evaluate features, as well as rules. Following predictive success of a rule, the probability of sampling the features which participate in the rule increases; following a misprediction, the probability of sampling participant features decreases. Thus features which participate in successful rules increase in salience, or probability of sampling. Features which participate in unsuccessful rules decrease in salience.

Focused sampling should facilitate learning rules in certain structures. When the input affords multiple rules among an overlapping set of features,

---

2 The sequence is probably more complicated. A distinction probably needs to be drawn between some form of initial encoding of features and selection for relevance in rule learning. For example, selection of features used in rule learning might be influenced by the observed value of the feature. This problem was noted in Fisher and Zeaman's learning model, and they embed a feedback loop in the selection process to accommodate it. The complication is not addressed in the present simulation, but would not form a problem in principle.

3 The salience of a feature is simply its probability of being sampled and need not imply perceptual prominence. Nonperceptual features could have saliences in this terminology. Rule "strength" and feature "salience" are mnemonic terms and are used in preference to calling either value a "weight."
discovery of any one initial regularity will facilitate discovery of other rules or regularities which share features with the initial rule. In effect, focused sampling is a bias to assume that regularities in input will not be distributed uniformly among features; rather, focused sampling assumes that if a feature is predictive in one rule, then the feature is likely to participate in other successful rules. When this bias is appropriate, it facilitates rule learning. A particular rule will be learned faster when it is part of a system of rules among an overlapping set of features than when that same rule occurs in isolation. This effect is called *clustered feature facilitation*. An individual rule,

\[
\text{Rule}_{1a}: \{f_x = N \rightarrow f_x = A\}
\]

is learned faster when its features participate in other rules, such as Rule\(_2\) and Rule\(_3\) below,

\[
\begin{align*}
\text{Rule}_{1a}: \{f_x = N \rightarrow f_x = A\} & \quad \text{Rule}_{1b}: \{f_x = A \rightarrow f_x = N\} \\
\text{Rule}_{2a}: \{f_x = J \rightarrow f_x = A\} & \quad \text{Rule}_{2b}: \{f_x = A \rightarrow f_x = J\} \\
\text{Rule}_{3a}: \{f_x = A \rightarrow f_x = J\} & \quad \text{Rule}_{3b}: \{f_x = J \rightarrow f_x = N\}
\end{align*}
\]

than when Rule\(_{1a}\) occurs in isolation. In contrast, in a simple model with only rule learning and no attentional learning, the learning of an individual rule would not benefit from the presence of multiple rules among the same features; rather, learning about one rule would be independent of learning about others.

Focused sampling benefits learning only when the structure to be learned is consistent with the bias that focused sampling assumes. It does not benefit rule learning if the structure of input does not afford intercorrelated features. If regularities are scattered through the system, predictions that a feature will be used in additional rules because it is used in one will not be sound. However, psychologically natural categories typically provide the correlational coherence assumed by focused sampling. Within a domain, if a feature participates in one regularity it is likely to participate in others. For example, singing predicts having feathers, flying, and living in trees. Instances picked out by one predictive rule are much the same set as instances picked out by other rules in the cluster (Darwin, 1962). For a learner equipped with focused sampling, noticing that many singing creatures fly would increase attention to the characteristics involved—singing, and perhaps vocalization as well, depending on the representation. This in turn increases the chance of noticing, say, that singing creatures live in trees.4 Discovery of

---

4 One desirable property of this data-driven inductive procedure is that it may dovetail with theory-based learning. Correlated features discovered by data-driven processes are good targets to be linked by causal relations and identify likely loci for theory-based projection. Features identified as important from theory-based projections may be useful loci for looking for other regularities in the data. Thus data-driven focused sampling and theory-driven learning may constructively feed off one another because of the importance of correlational structure for both.
such relations (and any causal basis) might be the basis for establishing a novel category or for transforming an initial similarity-based grouping to a more analyzed form. Shipley, 1988b) propose a transition from similarity-based class to predicate-based category. In established, adult categories, information about feature co-occurrence is a central part of both formal and informal concepts of natural kinds. If the information is important in initial learning as well, focused sampling provides a mechanism for making such discoveries more efficiently.

Syntactic categories of natural language also afford coherence among multiple criteria. Maratsos and Chalkley (1980) have suggested that syntactic categories may be learned by discovery of such coherence. The internal feedback model with focused sampling provides a mechanism for making such learning more effective. For example, in the German gender system, the various cases of definite and indefinite articles, adjective endings, and phonological form of the word are all interpredictive. A distributional learner equipped with focused sampling would capitalize on this correlational coherence to facilitate learning the gender classes. Prior work on learning syntactic categories in miniature artificial languages found evidence for clustered feature facilitation (Billman, 1983; Billman, 1988; Billman, Heit, & Dorfman, 1987).

The focused sampling plan extends the Zeaman and House attentional learning model to address learning without feedback. However, this transition is fundamental, since one of the key predictions changes. In the Zeaman and House model, predictor features can only compete with one another. In the focused sampling model, covarying predictors produce mutual facilitation, as well as the possibility for competition. On the one hand, covariation with and prediction of other features is the basis for increasing each feature’s salience or strength. On the other hand, features must also divide up limited attentional or processing resources. The effects among covarying features are more complex in learning without feedback because any feature may be either the predicted or predictor, in different rules; all features have the same dual status. In learning with feedback, any feature serves in only one role so the feedback feature has a different formal status than the predictor features. Put differently, learning with feedback is a special case where the learner is told which feature to predict (the feedback) and which features to treat as predictors (the cues). This dramatically reduces the possibilities the learner needs to consider to succeed in the task. Where the task is specified this way, learners may indeed narrow down their learning to discovery of some predictor sufficient to predict the feedback.

CARI: A Simulation of the Focused Sampling Plan
Category And Rule Inducer (CARI) simulates the internal feedback model for rule learning. It learns contingencies between features, or simple condi-
tional rules, and uses the focused sampling plan to direct learning toward the more valuable, predictive features. By running the model with and without the focused sampling component, for a variety of learning sets, we can evaluate the effects of focused sampling.

Given a set of schematized objects, CAR1 learns conditional rules between features with rule strengths approaching conditional probability. CAR1 also maintains a salience rating for each feature which determines its probability of being sampled in rule learning. CAR1 can run either with independent sampling, where saliences are fixed, or with changeable saliences as specified by focused sampling.

**Representation.** CAR1 represents rules, rule strengths, features, and feature strengths. CAR1 learns conditional rules about co-occurring features, of the form

\[ R_m = \{f_i = V \rightarrow f_j = W\} \]

where \( m \) indexes the rule, \( i \) and \( j \) index the two different features in the rule, and \( V \) and \( W \) index the feature values. This rule can be read as "feature \( j \) has value \( W \) given that feature \( i \) has value \( V \)." For every possible rule \( R_m \), CAR1 computes and stores a strength value, \( St(R_m) \), which estimates at time \( t \) the validity of the relationship between the two values of the features specified in rule \( R_m \). The strength is a number between 0 and 1 which approximates the conditional probability of the rule, as estimated from input data. We compare learning curves of \( St(R_m) \) generated in different conditions but with the true conditional probability the same across conditions. For example, a faster rise of \( St(R_m) \) in one condition over another shows facilitation for learning in the first condition.

For each feature \( f_i \), CAR1 maintains a salience rating \( Sal(f_i) \) which ranges from 0 to 1 and represents the importance of the feature at the current time \( t \). Feature salience or weight need not be fixed by perceptual factors and may be influenced by learning. Increasing \( Sal(f_i) \) increases the probability of sampling \( f_i \) by the rule-learning procedure. To convert salience to probability of sampling, salience is normalized over all possible features, as in Lute's decision rule (Lute, 1959). For example, if there are four features each with salience 0.5, then the probability of sampling a particular feature as a predictor is 0.025. In the following experiments, CAR1 starts with all features equally salient; if a particular task were being modeled, feature saliences could be set to different values initially.

**Revision Formulae for Rule and Attentional Learning.** The learning formula for revising the strength of rules is the delta rule shown here. Rules of this general form are used in many learning models (Rescorla, 1972; Rumel-
A\(\Delta St(R) = \alpha[T - St(R)]\), where \(\alpha\) = rate parameter for rule learning and \(\alpha > 0\)

\(T = 1\) following successful trials and

\(T = 0\) otherwise;

\(St_{t+1}(R) = St_t(R) + \Delta St_t(R)\).

Following successful predictions, the delta rule increases rule strength proportional to the difference between current and maximum strength. Following failures, strength is decremented proportional to the difference between current and minimum strength. The incremental nature of the rule is important but our simulation is not designed to vary and test the particular form of the rule. The wide use of the rule suggests it is robust and appropriate over a wide class of learning problems. The approximation by CARI takes place over many calls to the program, and initially all \(St(R_n)\) are set to a uniform, low value.

Salience is also changed by a delta rule:

\(\Delta Sa(f_i) = \beta[T - Sa(f_i)]\), where \(\beta\) = rate parameter for attentional learning

\(T = 1\) following successful trials and

\(T = 0\) otherwise.

\(Sa_{t+1}(f_i) = Sa_t(f_i) + \Delta Sa_t(f_i)\).

Independent sampling is simulated by keeping saliences fixed; if \(\beta = 0\) no attentional learning occurs and the probability of sampling a feature will not change with experience. For values \(0 < \beta < 1\) salience changes with learning. Changing the salience of one feature does not affect the salience of other features directly. However, probability of sampling is indirectly affected. Our simulation assumes a fixed, not just limited attentional capacity. Therefore, if \(Sa(f_i)\) decreases while \(Sa(f_2)\) and \(Sa(f_3)\) remain the same, \(f_2\) and \(f_3\) will be selected with greater probability when competing with \(f_1\).

**The NOTE Procedure.** NOTE is the procedure which learns from observing one object. It is called on a list of schematized objects and applies once to each object to make one learning trial. The order of object selection within a trial is determined randomly. Learning is evaluated after a block of trials. Each stimulus object is represented as a list of feature values, such as \(1 1 3 4\).

NOTE performs the following procedure on each object one at a time. It generates a rule, tests the rule on the current object, and changes strengths and saliences accordingly.

1. **Choose predictor feature.** One feature, \(f_i\), is sampled from all possible features. The sampling probability of each feature is determined by
2. **Choose predicted feature.** A second feature, \( f_j \), is selected from the remaining features, again using saliences to determine probability of sampling. The predicted feature is chosen from all features except \( f_i \).

3. **Observe value of predictor feature.** NOTE finds value \( v \) for the rule to be tested by looking at \( f_i \) of the current object and taking the value of this feature.

4. **Project value for predicted feature.** The predicted value \( w \) is chosen probabilistically from the strengths of all rules \( \{ f_i = v \Rightarrow f_j = W \} \), where \( W \) indicates any possible value of \( f_j \). For example, if there are four possible values of \( W \) to select, associated with four rules \( R_1, R_2, R_3, \) and \( R_4 \), and \( St(R_1) = St(R_2) = St(R_3) = 0.05 \) while \( St(R_4) = 0.15 \), then there is a probability of 0.5 for using \( R_4 \).

5. **Test rule and revise strengths.** The rule \( \{ f_i = v \Rightarrow f_j = w \} \) is tested against the data, the current object. Two types of learning take place, modifying rule strength and feature salience. Both types depend on agreement between hypothesis and data.
   a. **Adjust rule strength.** If NOTE's prediction of value \( w \) was correct, then the strength of this rule is increased, making \( w \) more likely to be predicted in the same situation. If the prediction was wrong, rule strength is decreased.
   b. **Adjust saliences.** If the prediction was correct, the saliences of \( f_i \) and \( f_j \) are increased, making these features more likely to be sampled in the future. If the prediction was wrong, the saliences of \( f_i \) and \( f_j \) are decreased.

Thus, a feature's performance in one rule affects its use in another. Learning about one aspect of the environment affects learning about other aspects.

**Parameters.** CARI has four parameters, three of which were fixed for these experiments; the fourth varied to show application of focused sampling. Feature salience for each feature, \( S_a \), was set to a value of 0.125 before any data was seen. Piloting demonstrated that lower values of \( S_a \) led to great variability among subjects. Similarly, strength of each rule, \( St \), was given an initial value of 0.01. This value kept the variability of our learning measures fairly low. The learning rate parameter for rule strengths, \( \alpha = 0.02 \), controls how fast CARI will change its estimates of conditional probability. As shown by Widrow and Stearns (1985), values of \( \alpha \) which are too high lead to poorer approximations of conditional probability because the asymptotic variance of the estimate is proportional to \( \alpha \). Smaller values of \( \alpha \) lead to a more reliable estimate, but learning takes place at a slower rate. The parameters \( S_{t_{0}}, S_a \) and \( \alpha \) were chosen to balance performance and varia-
bility but were not varied experimentally here. During exploratory piloting with roughly two dozen runs with different parameters, we varied $S_0$ and $S_a$ from 0.001 to 0.5, and $\alpha$ from 0.001 to 0.2, sampling unsystematically among parameter combinations. We used simple stimuli for most of this exploration. Over this range we consistently got benefit from focused sampling, with the amount and reliability varying considerably. Hence, while the values used in the experiments were set rather arbitrarily, the existence of the primary phenomenon, an advantage with focused sampling, is not dependent on a very restricted range of parameter values.

The experimental parameter, $\beta$, controls the rate of changing saliences. When simulating independent sampling, $\beta$ equals zero. Then saliences remain the same regardless of the nature of the learning set and there is no attentional learning. With values of $\beta$ between 0 and 1 exclusive, saliences change as required for focused sampling. With a setting greater than zero, attention is directed to features which participate in structured relations with other features. Although $\alpha$ was not varied systematically, informal exploration suggested that changes in $\beta$ do not interact dramatically with value of $\alpha$.

**General Simulation Method.** The simulations were done as experiments, where conditions varied attentional learning, stimuli configuration, or the presence of feedback. Subjects within a condition varied due to probabilistic selection of objects, features, and values. Subjects in contrasting conditions were matched or yoked so that the same "subject" in different conditions was always exposed to the same objects in the same order. Experiments used 16 of these simulated subjects per condition.

All stimulus sets provided a target set of rules among correlated features. In all stimulus sets, at least the first two features were correlated; schematically, items were as follows: (11x--x, 11x--x, ..., 22x--x, 22x--x,...) where -- indicates possible elided features and . . . indicates possible additional items. The target rules were the four symmetrical rules between the first two features: \{f_1 = 1 \rightarrow f_1 = 1\}, \{f_1 = 1 \rightarrow f_1 = 1\}, \{f_1 = 2 \rightarrow f_1 = 2\} and \{f_2 = 1 \rightarrow f_1 = 2\}. Strengths of these four rules were averaged and used as the measure of how well subjects learned. Since all four rules are based on the same two features and all have the same amount of evidence, averaging the four rules provides a more stable estimate of rule strength. Conditions differed in the availability of additional rules beyond these four, and the ef-

---

---

1 Equivalently, matching across conditions could be viewed as using a yoked subject design. Subject matching is used in analyses where learning with and without focused sampling is calculated for each set of data from the same input sequence. Throughout, reference to the "same subject" or to yoked subjects refers to sets of data from the same input sequence. Using the analogy to real subjects facilitates presentation. Statistically, the main effect is using difference scores for each subject or paired comparisons with their reduced degrees of freedom.
ffects of certain additional rules on learning the target rules are the focus of Experiments 3–6.

Three groups of experiments were run. The first tested for the predicted facilitation from focused sampling, and investigated the effects of varying the $\beta$ parameter and of the presence of exceptions to the target rules. The effects of focused sampling were evaluated by running the simulation with and without the focused sampling procedure. The purpose of this group of experiments was to establish the benefit from focused sampling over a range of learning stimuli. The second group of experiments examined the interaction of the benefit from focused sampling with structure of the stimuli. Experiments varied the number and distribution of covarying (relevant) and unsystematic (irrelevant) features, and the number of objects. These experiments are analogous to experiments on concept learning with feedback which varied the number of relevant and irrelevant cues (Zeaman, Campione, & Allen, 1970). The third group of experiments compared learning with and without feedback to explore the different roles attentional learning plays in the two circumstances.

DEMONSTRATIONS OF BENEFIT: EXPERIMENTS 1 & 2

Experiment 1 looked for an advantage of focused sampling on learning conditional rules available in a stimulus set. It varied values of $\beta$, the attentional learning parameter, and identified the value to be used for subsequent experiments. Piloting led us to expect that very low $\beta$s would give little benefit, moderate $\beta$s would lead to significant benefit, and higher $\beta$s would add much more variability to the learning measures of different subjects.

Experiment 2 tested for advantage from focused sampling when learning from more complex stimulus sets with exceptions; an advantage here is important if focused sampling is to contribute to inductive learning of naturally occurring regularities. A measure of the effect of focused sampling is introduced, the focused sampling advantage. The focused sampling advantage is the difference in rule strength following learning with or without focused sampling, assessed at a criterion point when both conditions receive equal exposure to the data. A large positive difference indicates a large advantage for focused sampling.

Experiment 1

Method. The stimulus set had four objects, each with four features; schematically, the objects were 1111, 1122, 2233, 2244. Only the first two features covaried across objects. Six conditions were run, with $\beta$ set to 0, .001, .002, .02, .2, and .4. Performance was evaluated on two measures after
each block of 50 trials: 1) average rule strength of the four target rules and 2) average salience of the two target features.

**Results.** Experiment 1 showed that focused sampling can facilitate learning conditional rules from a simple stimulus set. For this wide range of $\beta$ parameter values, the higher the attentional learning parameter, the greater the rule strength. This was consistent throughout the learning curve. Figure 1A shows the learning curves for rule learning when $\beta = 0$ and $\beta = .02$. We compared learning using each rate of focused sampling to learning using independent sampling. Comparison was made after a criterion number of trials (250), set when mean strength of target rules in the independent sampling condition was equal to $50 \pm 2$. This learning criterion was selected to fall in a relatively stable part of the learning curve. First, across levels of $\beta$ from 0 to .4, $\beta$ was a significant factor on amount of learning (Kruskal-Wallis statistic = 15.44, $p < 0.01$, inhomogeneous variance). Each focused sampling condition was compared to the independent sampling condition, using paired $t$-tests in matched or "within subject" analysis. Across three orders of magnitude (.001 to .4), rule strength increased.

![Learning curves for rule strength](image)

**Figure 1A.** Results in Experiment 1: Learning curves for rule strength.
with each increase in \( \beta \), but variance also increased. With higher \( \beta \), faster change of salience may lead to greater strengths for target rules. However, an early idiosyncratic success in predicting irrelevant features can also draw attention sharply away from relevant features and target rules. Indeed, for three subjects with \( \beta = 0.4 \), target rule strength was negligible even in the last block run. An intermediate value of \( \beta = 0.02 \) produced large benefit of focused sampling and modest variance to result in the most reliable benefit of focused sampling (\( t (15) = 2.72, p < 0.01 \)).

The relation of attentional learning and rule strength can be seen by comparing Figures 1A and 1B. A sharp initial rise in salience of relevant features precedes the points of greatest advantage for rule learning. With continued learning, salience of relevant features decreases. As learning progresses, knowledge about the co-occurrence of idiosyncratic as well as systematic features is acquired. When the learner knows the idiosyncratic, instance-level structure too, the learner can make successful predictions from all features, so the target features lose their relative advantage.

![Attentional Learning Curves](image)

**Figure 1B.** Results Experiment 1: Learning curves for salience of target features. Figures 1A and 1B show data for learning with and without focused sampling. Each point is based on 16 subjects. Bars indicate standard error.
Experiment 2

Method. The experiment had four conditions, varying whether focused sampling was used and whether the rules had exceptions. The learning stimuli in all four conditions consisted of 16 objects. Each object had 8 features. Three of the eight features were correlated with one another. Target rules assessed are the same four as in Experiment 1, namely \( \{f_1 = 1 \rightarrow f_2 = 1\}, \{f_1 = 1 \rightarrow f_1 = 1\}, \{f_1 = 2 \rightarrow f_2 = 2\} \) and \( \{f_2 = 2 \rightarrow f_1 = 2\}\). In the two conditions with exceptions, all rules had violations. Table 2 shows the learning stimuli.

In all four conditions, 10 blocks of 75 trials were run. The learning curves in Figure 2 show the increase in rule strength with and without focused sampling (\( \beta = 0 \) and \( \beta = 0.02 \)), for stimulus sets with and without exceptions. The learning criterion was defined to be the number of trials needed without focused sampling for the average performance measures to be 50 ± 2. The number of trials to criterion was determined for the no-exception condition and for the exception condition. These conditions were then run with focused sampling for a matched number of trials to compare learning following equivalent exposure to input. The benefit of focused sampling for each condition was the difference between learning measures with focused sampling and without.

Results. The learning curves in Figure 2 show the increase in rule strength with and without focused sampling (\( \beta = 0 \) and \( \beta = 0.02 \)), for stimulus sets with and without exceptions.

The focused sampling advantages for all 16 subjects and both stimulus sets were positive. The mean advantage was 26.18 for the stimuli without any exceptions and 11.71 for the stimulus set with exceptions; both scores reflect an advantage much greater than a chance score of zero (with exceptions \( t(15) = 18.38; \) no exceptions \( t(15) = 9.12; \) \( p's < .0001 \)). The advantage
TABLE 2
EXPERIMENT 2
Learning Stimuli

<table>
<thead>
<tr>
<th>Without exceptions</th>
<th>With exceptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1111111111</td>
<td>*1221111111</td>
</tr>
<tr>
<td>1112222222</td>
<td>*1122222222</td>
</tr>
<tr>
<td>2223333333</td>
<td>*2113333333</td>
</tr>
<tr>
<td>2224444444</td>
<td>*2214444444</td>
</tr>
<tr>
<td>1111234443</td>
<td>1111234443</td>
</tr>
<tr>
<td>111234134</td>
<td>111234134</td>
</tr>
<tr>
<td>222341221</td>
<td>222341221</td>
</tr>
<tr>
<td>222412312</td>
<td>222412312</td>
</tr>
<tr>
<td>111334411</td>
<td>111334411</td>
</tr>
<tr>
<td>111443322</td>
<td>111443322</td>
</tr>
<tr>
<td>222112233</td>
<td>222112233</td>
</tr>
<tr>
<td>222221144</td>
<td>222221144</td>
</tr>
<tr>
<td>111432143</td>
<td>111432143</td>
</tr>
<tr>
<td>111341234</td>
<td>111341234</td>
</tr>
<tr>
<td>222123412</td>
<td>222123412</td>
</tr>
<tr>
<td>222234123</td>
<td>222234123</td>
</tr>
</tbody>
</table>

* Exceptions. The feature values of the first three features in these items disrupt the dominant correlational pattern. For example, the first and third items violate the standard feature assignments for the first two features, with values of 12 and 21 rather than the standard values of 11 and 22.

LEARNING CURVES FOR STRENGTH OF TARGET RULE
WITH AND WITHOUT FOCUSED SAMPLING

Figure 2. Results Experiment 2: Learning with and without exceptions. The separation of the learning curves with focused sampling from the curves without focused sampling reflects the focused sampling advantage.
was greater for learning without exceptions than for learning from the stimulus set with exceptions ($p < 0.0001$, $t(15) = 6.29$). These difference scores are larger here than the advantage of 6.35 at the same criterion in Experiment 1.

**Discussion: Experiments 1 & 2**
Experiments 1 and 2 demonstrated that use of focused sampling leads to faster learning of target rules. As well as increasing rate of rule learning, focused sampling also increases variability. It benefits learning both for stimulus sets with a small number of objects with a few features each and for stimulus sets with more objects, more features, and more feature correlations. Comparison between Experiment 1 and 2 (at criterion) shows that focused sampling provides more benefit on more complex stimulus sets; difference scores and significance levels were much greater in Experiment 2 for learning without exceptions than in Experiment 1. In addition, the benefit is greater for the no-exception condition than the condition with exceptions. Since focused sampling is tuned to discovery of regularity, it helps learning more when the data provide more regularity. Lower asymptotic rule strength in the exception condition may also be responsible for the smaller amount of focused sampling advantage here.

**EFFECTS OF STIMULI STRUCTURE: EXPERIMENTS 3, 4, 5, & 6**
Experiments 3, 4, 5, and 6 investigated how the organization of systematic and unsystematic variability among features affects the amount of benefit from focused sampling. We varied number of objects, number of features, and distribution of features into those which covary systematically and those which are idiosyncratic. We refer to features which systematically covary with others as relevant and features which vary idiosyncratically as irrelevant. The total number, ratio of relevant to irrelevant, total relevant, and total irrelevant features are mutually constraining, so one cannot be varied independently of all the others. Further, we expected that effects would interact. Experiment 3 varied the number of objects and the mix of relevant and irrelevant features, while total number of features was held constant at eight. Experiment 4 varied the number of irrelevant features, at two levels of relevant features, and Experiment 5 varied the number of relevant features at two levels of irrelevant features. Experiment 6 varied the total number of features, keeping the ratio of relevant to irrelevant constant. Specific predictions are presented with each experiment.

**Experiment 3**
First, we expected that focused sampling would be more beneficial when there are more objects, given equivalent total exposure. A larger sample of
objects would allow better separation between the reliable target rules and the spurious correlations true of only one object. An idiosyncratic correlation will be more likely to be contradicted as the sample size increases. When the rules among the relevant features have relatively greater predictive merit, the irrelevant features will gain less attention, so the rules among relevant features will have less competition.

Second, we predicted clustered feature facilitation, namely, that if additional features covary with the target features, the benefit of focused sampling would increase. If a feature participates in multiple, well-supported rules, then that feature would have its salience increased more rapidly than a feature participating in fewer rules.

**Method.** Stimuli sets varied in number of objects and number of relevant features. Four or eight objects per set were used. The number of relevant features varied from 2 to 7. Thus, twelve sets of stimuli were used. The objects in all sets had 8 features. Table 3 shows the learning stimuli. Focused sampling advantage was computed at the criterion number of trials, as in Experiments 1 and 2.

**Results.** Figure 3 shows the amount of advantage of focused sampling, with 4 and 8 objects and varying assignment of the eight features as relevant or irrelevant. An analysis of variance was done with number of objects and

<table>
<thead>
<tr>
<th>Ratio of Relevant to Irrelevant features</th>
<th>4 Objects</th>
<th>8 Objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>2/6</td>
<td>11111111 11222222 22333333 22444444</td>
<td>11111111 11222222 22333333 22444444</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11123443 11234134 22341221 22412312</td>
</tr>
<tr>
<td>3/5</td>
<td>11111111 11222222 22333333 22444444</td>
<td>11111111 11222222 22333333 22444444</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11112344 11123413 22234122 22241231</td>
</tr>
<tr>
<td>4/4</td>
<td>11111111 11112222 22223333 22224444</td>
<td>11111111 11122222 22223333 22224444</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11111234 11112341 22223412 22224123</td>
</tr>
<tr>
<td>5/3</td>
<td>11111111 11111222 22222333 22222444</td>
<td>11111111 11111222 22222333 22222444</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11111123 11111234 22222341 22222412</td>
</tr>
<tr>
<td>6/2</td>
<td>11111111 11111122 22222233 22222244</td>
<td>11111111 11111122 22222233 22222244</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11111112 11111123 22222234 22222241</td>
</tr>
<tr>
<td>7/1</td>
<td>11111111 11111112 22222223 22222224</td>
<td>11111111 11111112 22222223 22222224</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11111113 11111114 22222221 22222222</td>
</tr>
</tbody>
</table>
assignment of relevant and irrelevant features as factors. Both main effects and interaction were significant (objects $F(1,180) = 94.00$; relevant features $F(5,180) = 12.90$; interaction $F(5,180) = 6.17$; all $p$'s < .001). Table 4 shows cell means and post hoc comparisons; a difference of 8.52 is required for cell means to differ significantly at the .05 level (Tukey's HSD). Focused sampling provided much greater benefit when the learning set had eight objects, rather than four. Pairwise comparisons showed an advantage for eight objects when 2, 3, 4, or 5 features were relevant. For each of these four contrasts, focused sampling provided more benefit from the 8-object set. The number of relevant features had a strong but nonlinear effect on focused sampling advantage; the focused sampling advantage increased to a maximum at half relevant, half irrelevant and then decreased with a higher proportion of relevant features.

Perhaps the most dramatic finding was the performance with four objects at 2 relevant features. Here, focused sampling resulted in significant cost to learning, not benefit ($t(15) = 2.64$, $p < .01$).

Discussion. The results strongly indicate that focused sampling provides greater benefit for sets with more objects, so long as those objects decrease
the reliability of rules involving the irrelevant features. The relative reliability of target rules compared to idiosyncratic patterns was greater in the 8-object rather than the 4-object set; this contrast holds generally for larger versus smaller object sets. With a small sample of objects which belong to the same category, there will be many local or idiosyncratic patterns which are consistent within that sample. Though it may not generally be true that doctors with mustaches wear turtlenecks, it might coincidentally be true in a small sample; as the sample size increases it becomes less likely that such predictive "rules" would have consistent support. As sample size increases, relative competition from idiosyncratic rules decreases. 6 With 8 objects and 3 of 8 features relevant, the focused sampling advantage was 18.90. This value can be compared to learning in Experiment 2 with 16 objects and 3 of 8 features relevant where the focused sampling advantage was 26.18.

The focused sampling advantage increases with more objects. Indeed, with a small set of objects in which the most reliable structure is at a very specific, instance level, focused sampling can impair learning a more general rule. We predicted that a larger and more varied set of instances would increase the value of focused sampling on learning a target rule but we had not anticipated cost with small numbers of examples. This is an interesting characteristic of focused sampling revealed by the simulation. The system may allocate attention to individual instance patterns or to general ones, depending on the relative reliability and amount of evidence available for either sort.

The distribution of the 8 features into relevant and irrelevant cues also had a major effect on the amount of benefit. On the one hand, increasing

---

6 Consider the situation with eight objects and two relevant features. In the 8-object set the target rules, such as \{f_1 = 1 \Rightarrow f_2 = 1\}, were completely reliable, and only the two target features participated in any exceptionless rules. The rules between two irrelevant features or between a relevant and irrelevant feature had exceptions. Therefore, the rules in which the relevant features participated were rewarded most consistently. In addition, any rule about the relevant features applied more often (to 4 of 8 objects) than a rule about the less frequent feature values of the irrelevant features (only 2 of 8 objects have \(f_3 = 2\)). Both frequency and reliability favored the relevant features.

The situation is different when only 4 objects were used. For just four objects, there was only slightly more evidence for a rule true of 2 of 4 objects (such as \(f_1 = 1 \Rightarrow f_2 = 1\)) than for a "rule" true of 1 of 4 objects (such as \(f_3 = 1 \Rightarrow f_4 = 1\)). There is little difference in the evidence supporting a more general rule and a rule at the level of one individual instance when there are few instances. Further, rule reliability worked against the redundant, target features. "Rules" among irrelevant features were perfectly reliable. Take the case where there are two redundant features. Given the value of any of the 6 irrelevant features, the values of any of the remaining 7 features is known, since any of these values uniquely identifies an object. The relevant features were not so predictive: given that \(f_1 = 1\), \(f_3\) might equal 1 or 2. Thus, in the 4-object set, focused sampling rewards the features which differ with each instance, features 3 to 8, since these are even more reliable than the relevant features. While each rule about features 3 to 8 applies less often, the values are always predictable. Focused sampling will allocate attention to the most predictive features, which need not be the designated target features.
the number of relevant features should increase the estimated predictivity of target features, increase attention to those features, and thus produce more advantage from focused sampling. On the other hand, reducing the number of irrelevant features reduces the pool of features from which relevant features can steal attention. This produced the strong, nonmonotonic effect of varying the number of the eight total features which were relevant. Maximum benefit was found when about half the features are relevant and half irrelevant.

Experiment 4

Experiment 4 investigated the effects of the number of irrelevant features, at two levels of relevant features. Unlike Experiment 3, the total number of features increased as the number of irrelevant features increased. Some irrelevant information is necessary for focused sampling to be beneficial, since there must be some irrelevant features from which the relevant features can steal attention. Given a fixed attentional capacity, if some features are sampled more frequently, other features must be sampled less. Thus, to some extent adding irrelevant features should increase the benefit of focused sampling. This effect may not continue indefinitely. For a fixed number of instances, adding irrelevant features also adds unimportant, idiosyncratic
structure which nonetheless draws attention. For a fixed number of learning instances, the benefit of focused sampling should initially increase and then attenuate or decrease as irrelevant features are added.

**Method.** Experiment 4 varied number of irrelevant features from one to six, while keeping the relevant features at either two or three features. Table 5 shows the stimuli design.

The procedure is the same as in Experiment 3.

**Results.** An analysis of variance showed that number of irrelevant features, $F(5,180) = 10.40$, number of relevant features, $F(1,180) = 12.71$, and their interaction, $F(5,180)$, were all highly significant, $p < .0001$. Data is shown in Table 6 and Figure 4. Consistent with the hypothesized facilitation from clustered features, greater advantage came from more relevant cues.

---

**TABLE 5**
**EXPERIMENT 4**
Learning Stimuli

<table>
<thead>
<tr>
<th>Number of Irrelevant Features</th>
<th>2 Relevant Features</th>
<th>3 Relevant Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  11x...22x</td>
<td>111x...222x</td>
<td></td>
</tr>
<tr>
<td>2  11xx...22xx</td>
<td>111xx...222xx</td>
<td></td>
</tr>
<tr>
<td>3  11xxx...22xxx</td>
<td>111xxx...222xxx</td>
<td></td>
</tr>
<tr>
<td>4  11xxxx...22xxxx</td>
<td>111xxxx...222xxxx</td>
<td></td>
</tr>
<tr>
<td>5  11xxxxx...22xxxxx</td>
<td>111xxxxx...222xxxxx</td>
<td></td>
</tr>
<tr>
<td>6  11111111 11222222 22333333 22444444</td>
<td>11111111 11222222 22333333 22444444</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11123443 11234134 22341221 22412312</td>
<td>11123443 11234134 22334122 222412312</td>
</tr>
</tbody>
</table>

**TABLE 6**
**EXPERIMENT 4**
Focused Sampling Advantage
Effect of Irrelevant Features at Two Levels of Relevant Features
(8 Objects, 16 Subjects Per Cell)

<table>
<thead>
<tr>
<th>Number of Irrelevant Features</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 relevant mean</td>
<td>10.38</td>
<td>14.26</td>
<td>17.94</td>
<td>16.86</td>
<td>12.41</td>
<td>8.04</td>
</tr>
<tr>
<td>(st dev)</td>
<td>(6.93)</td>
<td>(7.00)</td>
<td>(5.14)</td>
<td>(7.28)</td>
<td>(6.28)</td>
<td>(7.51)</td>
</tr>
<tr>
<td>3 relevant mean</td>
<td>8.52</td>
<td>12.76</td>
<td>19.18</td>
<td>21.25</td>
<td>18.90</td>
<td>18.79</td>
</tr>
<tr>
<td>(st dev)</td>
<td>(6.46)</td>
<td>(6.51)</td>
<td>(6.81)</td>
<td>(4.45)</td>
<td>(4.61)</td>
<td>(6.03)</td>
</tr>
</tbody>
</table>

Advantage shown is difference in rule strength at criteria for focused and independent sampling.
Overall, three relevant features provided a benefit of 18.57, or about 137% of learning without focused sampling, while two relevant features provided a benefit of 13.31, or 127%. This advantage interacted with the number of irrelevant features. The effect of relevant features is substantial at larger numbers of irrelevant features, but makes little difference with three or fewer irrelevant features. As number of irrelevant features increases from one to three or four, advantage also increases from 9.45 to 18.51 or 19.06 respectively, but increase beyond this leads to decreasing or stable advantage. A Newman-Keuls post hoc comparison of cell means requires a difference of 7.3 for significance at the .05 level.

Expériment 5
Experiment 5 investigated the effect of relevant features. We predicted that, to a point, adding more relevant features that are correlated with the target features would increase the benefit of focused sampling by shifting more attention to the target features. However, eventually the addition of relevant features while holding irrelevant constant should attenuate or decrease the
advantage of focused sampling on a given target rule. At some number of relevant features, relevant features will compete primarily among themselves rather than increasing their strength primarily at the expense of the irrelevant features.

Experiment 5 varied number of relevant features from two to six, while setting the number of irrelevant features to either three or six. An analysis of variance with numbers of relevant and irrelevant features as factors showed that number of relevant features was a significant factor, $F(4,150) = 4.99$, $p = .001$, but the number of irrelevant features was not. In addition, the interaction was highly significant, $F(4,150) = 11.51$, $p < .001$. Figure 5 and Table 7 show the data. For 6 irrelevant features, the effect of relevant features is clear and as expected. Benefit increases from 2 to 4 relevant features and slightly diminishes as still more relevant feature are added. The increase from 2 to 3 relevant features is individually significant; Newman-Keuls requires a difference of 6.55 for individual cells to differ. For 3 irrelevant features, number of relevant features has less of an effect. An initial small increase in benefit from 2 to 3 features is followed by a small drop, and none of the five means differ significantly.

**Figure 5.** Results Experiment 5: Focused sampling advantage as a function of relevant features, where irrelevant features are held constant at 3 or 6 features.
TABLE 7
EXPERIMENT 5
Focused Sampling Advantage:
Effect of Relevant Features at Two Levels of Irrelevant Features
(8 Objects, 16 Subjects Per Cell)

<table>
<thead>
<tr>
<th>Number of Relevant Features</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 irrelevant mean</td>
<td>17.94</td>
<td>19.18</td>
<td>14.28</td>
<td>14.59</td>
<td>14.87</td>
</tr>
<tr>
<td>(st dev)</td>
<td>(5.14)</td>
<td>(6.81)</td>
<td>(2.93)</td>
<td>(5.82)</td>
<td>(7.64)</td>
</tr>
<tr>
<td>6 irrelevant mean</td>
<td>8.04</td>
<td>18.79</td>
<td>21.87</td>
<td>19.67</td>
<td>20.10</td>
</tr>
<tr>
<td>(st dev)</td>
<td>(7.51)</td>
<td>(6.03)</td>
<td>(5.18)</td>
<td>(3.80)</td>
<td>(5.91)</td>
</tr>
</tbody>
</table>

Advantage shown is difference in rule strength at criteria for focused and independent sampling.

This experiment suggests that when the total number of features is small or the proportion of relevant to irrelevant is large (perhaps 1/2 or more), then the number of relevant features will make little difference. Locating 4 relevant features among 3 irrelevant ones may not be much harder than locating 6 relevant features among 3 irrelevant. On the other hand, adding a few more relevant features to the target relation helps significantly if there are many irrelevant features.

Experiment 6
Experiment 6 tested the effect of total number of features, keeping relevant to irrelevant features at a constant one-to-one ratio. Five stimulus sets were used with 4, 6, 8, 10 and 12 total features. For each set, half the features were relevant, and half irrelevant. The procedure was the same as in Experiments 3, 4, and 5.

Results are shown in Figure 6. A one-way analysis of variance showed number of features to be a significant factor ($F(4,75) = 5.54, p < .001$). Larger numbers of features provided more benefit, with the largest increase from 4 to 6 features.

FOCUSED SAMPLING AND FEEDBACK:
EXPERIMENTS 7 & 8

CARI is a general-purpose system for learning any set of covariations between pairs of features. In the previous experiments, the NOTE procedure selected both the predictor and predictor features from any of the object's features. For all distinct features $f_i$ and $f_j$, CARI learned to predict the value
of $f_i$, given the value of $f_r$, and learned to predict $f_i$ from $f_j$ as well. We intended to model observational learning where there are no a priori restrictions on which contingencies are learned. In contrast, in learning with feedback, the feedback information has a privileged role: it is always the feature to be predicted. The task is defined more narrowly for the learner than in observational learning. To implement learning with feedback, the contingencies CAR1 learned were restricted to those predicting the one feature designated as the feedback. Contingencies between nonfeedback features were not learned, nor were predictions about the nonfeedback features, given the value of the feedback feature.

Thus, in some sense learning with feedback is just a special case of learning without feedback where only a subset of rules can be learned. But the results from Experiments 1 through 6 show facilitation between correlated features, something not predicted by models of learning with feedback and not obtained in feedback-bound learning experiments. Indeed, we expect that learning without feedback will lead to qualitatively different performance compared to learning with feedback. Experiments 7 and 8 used CAR1 to compare these two forms of learning directly.
To change CARI to implement learning from external feedback, the feedback feature is given a different status from the nonfeedback features. The NOTE procedure runs as follows:

1. Choose predictor feature from non-feedback features.
2. Choose predicted feature to always be the feedback feature.
3. Observe value of predictor feature.
4. Project value of the feedback feature.
5. Test rule and revise strengths and saliences.

Since the feedback feature is always chosen as the predicted feature, it is no longer given a salience. Selection of the predictor feature now uses the saliences of the nonfeedback features only. (Note that we refer to the feedback feature as part of the object; in the real world feedback may be obtained from a different source than the object itself, as when an experimenter provides a category label for a picture the subject views.)

The CARI model with feedback can be viewed as an extension of the Zeaman and House model; the major change is that features can take on multiple values. With feedback, CARI learns exclusively about predictors of feedback. We suspect that human learners are not so narrow and do incidentally learn other relations in input, even when feedback is provided. The CARI model with feedback is an extreme case of learning where predictions about the feedback feature are the only ones learned and predictions between nonfeedback features are ignored.

**Experiment 7**

How does feedback interact with focused sampling? Experiment 3 showed that without feedback, there is a nonlinear relationship between number of relevant features and amount of benefit from focused sampling. When there are few relevant features, adding another relevant feature leads to clustered feature facilitation and an increased benefit of focused sampling. Yet when there are many relevant features, adding another leads to increased competition between relevant features, and a decreased amount of benefit from focused sampling.

Since it directs attention to features which participate in successful predictions, focused sampling will certainly help learning about the relevant features even in learning with feedback. But since learning with feedback does not make use of covariation between nonfeedback features, we predicted that there would be no clustered feature facilitation. Rather than increasing with the initial addition of relevant features, the benefit of focused sampling should strictly decrease as the number of relevant features increases, due to competition between relevant features. Thus, we predicted that attentional learning with feedback will be qualitatively different in this respect from attentional learning without feedback.
Method. The 8-object stimuli set from Experiment 3 (see Table 3) was used, with number of relevant features ranging from 2 to 6. There was a no-feedback condition and a feedback condition. The no-feedback condition was a replication of the 8-object condition in Experiment 3, except that a slightly different performance measure was used: the average of the strengths of the rules \( \{ f_1 = 1 \Rightarrow f_1 = 1 \} \) and \( \{ f_1 = 2 \Rightarrow f_1 = 2 \} \) was used. That is, \( f_1 \) was treated as the feedback feature and only rules predicting the value of that one feature were assessed. For the feedback condition, CARI was changed as described above, so that it only learned to predict from nonfeedback features to the feedback feature. The first feature of each object was assigned to be the feedback feature. Since CARI did not learn rules of the form \( \{ f_1 = V \Rightarrow f_x = W \} \) in this condition, Experiment 7 could not use the rules \( \{ f_1 = 1 \Rightarrow f_1 = 1 \} \) and \( \{ f_1 = 2 \Rightarrow f_1 = 2 \} \) in the performance measure.

Exposure to input in the two conditions was matched in a manner similar to that in the previous experiments. For each condition, the simulation was run without focused sampling to determine the number of trials for that condition to reach the criterion of an average learning score of 50. Then the simulation was run again for the same number of trials but with focused sampling, to obtain difference scores representing the benefit of focused sampling for each subject. While the feedback and no-feedback conditions used different versions of CARI, focused sampling benefits can be compared between conditions because we equated degree of learning without focused sampling.

Results. Table 8 shows the average benefits of focused sampling. In the no-feedback condition, the benefit of focused sampling rises and then falls as the number of relevant features increases, but for the feedback condition,

<table>
<thead>
<tr>
<th>Mix of Relevant to Irrelevant Features (Rel-Irrel)</th>
<th>No Feedback</th>
<th>Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-6</td>
<td>mean 14.57</td>
<td>mean 20.37</td>
</tr>
<tr>
<td></td>
<td>st dev (7.94)</td>
<td>st dev (9.29)</td>
</tr>
<tr>
<td>3-5</td>
<td>mean 20.06</td>
<td>mean 16.48</td>
</tr>
<tr>
<td></td>
<td>st dev (7.08)</td>
<td>st dev (11.06)</td>
</tr>
<tr>
<td>4-4</td>
<td>mean 21.65</td>
<td>mean 12.92</td>
</tr>
<tr>
<td></td>
<td>st dev (5.68)</td>
<td>st dev (11.46)</td>
</tr>
<tr>
<td>5-3</td>
<td>mean 18.92</td>
<td>mean 10.82</td>
</tr>
<tr>
<td></td>
<td>st dev (5.07)</td>
<td>st dev (11.85)</td>
</tr>
<tr>
<td>6-2</td>
<td>mean 9.63</td>
<td>mean 7.86</td>
</tr>
<tr>
<td></td>
<td>st dev (7.24)</td>
<td>st dev (12.69)</td>
</tr>
</tbody>
</table>

Advantage shown is difference in rule strength at criteria for focused and independent sampling.
the benefit strictly decreases. To test the significance of these observations, for each condition a one-way ANOVA was performed, with number of relevant features as the independent variable and benefit of focused sampling as the dependent variable. For the no-feedback condition, trend analysis showed a significant linear component in the benefit scores, $t(75) = 2.09, p < 0.05$, and a highly significant quadratic component $t(75) = 5.42, p < 0.001$. For the feedback condition, the linear component was significant, $t(75) = 3.43, p < 0.001$, but there was no significant quadratic component, $t(75) = 0.31$.

**Experiment 8**

Experiment 7 showed evidence that the focused sampling advantage initially increases for additions of more relevant features during learning without feedback, but strictly decreases during learning with feedback. This result suggests that with feedback, the more a learner knows about one rule, the less he knows about another. Without feedback, the more a learner knows about one rule, the more he knows about other rules among the set of intercorrelated rules. The analysis for Experiment 7 used averaged performance scores; in Experiment 8 we looked for similar evidence at the level of individual rules learned by individual subjects. We expected that in the feedback condition, the strength of a subject's knowledge about a given rule would be negatively correlated with knowledge about another rule, due to attentional competition. However, in the no-feedback condition, we predicted that knowing one rule would facilitate learning about another related rule. Thus correlations between strengths of rules should be more positive in the no-feedback condition than in the feedback condition.

**Method.** The stimuli from Experiment 7 were used, except that the stimulus set with 2 relevant features could not be used because the performance measures required at least three relevant features. This experiment had a 2 by 4 design, with CARI version (feedback or no-feedback) crossed with number of relevant features (3 to 6). Feature 1 was always relevant and was used as the feedback feature in the feedback condition. We compared learning about rules of the form $\{f_2 = V \rightarrow f_1 = W\}$ and $\{f_3 = V \rightarrow f_1 = W\}$, that is, two rules predicting $f_1$ (the feedback feature in learning with feedback). Learning about conditional rules from $f_2$ to $f_1$ was measured by the average strengths of $\{f_2 = 1 \rightarrow f_1 = 1\}$ and $\{f_2 = 2 \rightarrow f_1 = 2\}$; learning about rules from $f_3$ to $f_1$ was measured by the average strengths of $\{f_3 = 1 \rightarrow f_1 = 1\}$ and $\{f_3 = 2 \rightarrow f_1 = 2\}$. The number of trials to criterion without focused sampling for each version of CARI was determined, then the simulations were run with focused sampling for the same number of trials. This experiment examined overall learning with focused sampling rather than difference scores showing the benefit of focused sampling.
TABLE 9
EXPERIMENT 8
Average Correlation between Strengths of Predictive Rules with and without Feedback for 8 Input Objects with 8 Features:
Ratio of Relevant to Irrelevant Features Varies
16 Subjects Per Cell

<table>
<thead>
<tr>
<th>Mix of Relevant to Irrelevant Features (Rel-Irrel)</th>
<th>3-5</th>
<th>4-4</th>
<th>5-3</th>
<th>6-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Feedback</td>
<td>0.106</td>
<td>0.010</td>
<td>0.237</td>
<td>0.492</td>
</tr>
<tr>
<td>With Feedback</td>
<td>-0.095</td>
<td>-0.369</td>
<td>-0.322</td>
<td>-0.337</td>
</tr>
</tbody>
</table>

Correlations shown are averages of the 16 individual, within-subject correlations in each condition.

Results. Table 9 shows the correlations between the \{f_2 = V \Rightarrow f_1 = W\} performance measure and the \{f_2 = V \Rightarrow f_1 = W\} performance measure. The correlation was computed over the 16 subjects in each cell of the experiment.

Note that correlations between strengths of the two target rules are all positive in the no-feedback conditions and all negative in the feedback conditions. Across the set of no-feedback conditions, the mean correlation was .211, whereas it was -.281 in the feedback conditions. These correlations were transformed to Z scores and a two-way ANOVA was performed. Version of CARI was a significant factor, \(F(1,3) = 53.62, p<0.01\); the effect of number of relevant features was not significant, \(F(3,3) = 6.12\). Thus, the correlations between rule strengths were significantly more positive in the no-feedback condition than in the feedback condition.

GENERAL DISCUSSION

Summary and Interpretation of Findings
Running a simulation, unlike running experiments with people, allows addition and deletion of a component process at will. Experiments typically allow direct control of stimuli but not the procedures subjects use. The ability to decompose the process in simulation allows a clear evaluation of what a target component contributes. By running a learning model with and without focused sampling, the effects of focused sampling on learning a variety of stimuli were evaluated. We found pervasive benefit from focused sampling, analyzed the effects of the correlational structure on the amount of focused sampling advantage, and investigated how feedback affects learning with focused sampling.

In all but one stimulus set, a simple and impoverished one (Experiment 3, 4 instances, 8 features total, 2 relevant features), focused sampling provided a trend toward facilitation, and usually a significant benefit. Its negative impact for the simple set was not anticipated.
The effects of stimuli structure on the amount of focused sampling advantage proved to be fairly complex. We summarize some major points. First, increasing the number of objects increased focused sampling advantage, by breaking up idiosyncratic competitor rules for which a small sample may provide consistent evidence. (Experiment 3, and informal comparison of analogous conditions in Experiments 2 and 3). Second, increasing the number of features per object, for a fixed ratio of relevant features, facilitated learning (Experiment 6).

Third, the distribution of relevant and irrelevant features greatly affected the amount of clustered feature facilitation. (Experiments 3, 4, 5). Adding interrelated, relevant features increased the benefit of focused sampling, particularly over small numbers of relevant features. These additional relevant features provided multiple cues that features participating in the target rule were predictive and should be explored, resulting in clustered feature facilitation. In contrast, adding relevant features decreased the benefit of focused sampling when roughly half or more of the features were relevant, due to increased competition. Benefit also increased with more irrelevant features from which attention could be reallocated.

Fourth, learning with and without feedback showed qualitative differences. Although the same basic learning mechanisms were used, redefining the task as being with or without feedback changed the information used to guide the application of the learning procedures. As a result, covarying predictors only competed with each other with feedback, but could facilitate as well as compete without feedback. Because of this facilitation, focused sampling also made a greater contribution without, rather than with, feedback when there were multiple relevant features. Experiment 7 showed a qualitative difference at a group level. Clustered feature facilitation was found only without feedback. Experiment 8 showed differences at the level of individual rules. With feedback, subjects' knowledge of one rule predicting the target feature was negatively correlated with knowledge of other rules; without feedback, subjects' knowledge of one rule was positively correlated with knowing other rules. Since most concept learning experiments have directed the subjects to attend to external feedback, it is not surprising that previous experimental results have shown competition and not facilitation between correlated features.

These findings lead to three very general suggestions. First, several factors which lead to increased task difficulty are associated with increased benefit from focused sampling; focused sampling helps more on hard problems. Second, structure of the whole system affects learning of components, even in our relatively simple learning situations. Third, the benefit from focused sampling depends on match between the organization of the input and the inductive bias toward correlational coherence which focused sampling assumes.
Evaluation of the Simulation

In evaluating a simulation, it is valuable to distinguish between the principles which the model is intended to explicate and the extra assumptions made for the implementation. Four fundamental principles of the model are incorporated in the simulation and deserve mention. First, focused sampling and the simulation assume that rules are learned from observation of instances, even in the absence of explicit feedback. The learner represents and uses information about what features are the important predictive ones, and what the relations are among these systematic features. Second, the learner is constrained in this learning process by limited processing capacity or attention. Thus, third, the learner needs an appropriate adaptive plan or inductive bias specifying which rules or hypotheses should be given priority. An absence of explicit external feedback imposes particularly great demands on the learner in structuring the learning problem.

Finally, one particular adaptive plan, focused sampling was incorporated. Any adaptive plan is tuned to some assumption about the nature of the material to be learned. Focused sampling assumes that natural categories reflect representable coherence in the world. By coherence we mean that regularities in feature co-occurrence are not distributed at random, but tend to form systems of relations among overlapping sets of features; in general, if a feature is predictive in one regularity, it will predict other features and participate in other rules. Focused sampling increasingly allocates attention or processing resources to features which participate in successful predictive rules.

The simulation added many assumptions about the details of rule representation and process of learning. Three are mentioned here because we do not believe that these assumptions, used by the simulation, are true of many classes of human category and rule learning. First, CARI's attentional limits are too severe. More plausibly, the learner is not restricted to noticing just one feature but samples from a distribution of numbers of features (as in earlier "multiple look" models such as Fisher & Zeaman, 1973). Second, unlike CARI, people are not restricted to learning rules about feature pairs but can relate multiple features, in at least some tasks (Anderson, Kline, & Beasely, 1979; Bruner, Goodnow, & Austin, 1956; Hayes-Roth & Hayes-Roth, 1977). Preference for instance-like "rules" which specify many features or more abstract rules among a few features presumably varies with task demands and stimuli characteristics. Third, CARI treats the predictor and predicted features either symmetrically, when learning without feedback, or completely asymmetrically for learning with feedback. A more general model should allow graded differences where (1) salience as a predictor and predictor can be differentially but continuously varied for each feature and (2) some features are more likely to be used as predicted features (e.g., intrinsically motivating outcomes or experimenter labels) and others
as predictors (e.g., those which are novel and surprising or those believed to play a causal role) without imposing a complete separation of roles. Presumably such graded roles for features is a better model for all but the most rigid or the most totally unconstrained learning situations.

No analytic arguments can be made demonstrating the scope of generality of the findings from the simulation. However, to the extent that benefit of focused sampling increases for more complex problems, the benefits demonstrated in the relatively simple cases explored here might be amplified when focused sampling is applied to increasingly realistic and complex problems.

Evaluation of Focused Sampling
We can evaluate focused sampling from a design perspective and from a psychological perspective. From a design perspective we consider whether focused sampling would make a strong contribution to inductive learning by a person or other entity. From a design perspective, focused sampling provides a simple, computationally feasible method of guiding observational learning and, to a first approximation, the contribution focused sampling makes increases with the complexity of the structure to be learned. Thus, focused sampling would be a valuable component in solving the induction problem for complex inductive learning tasks; further, it does not depend on explicit feedback or tutoring.

Earlier, we argued that constraint from specification of the representational language is insufficient. Even given a space of representable and potentially relevant distinctions, it is far from clear what the appropriate feature set might be. Definition or discovery of appropriate features is a problem for psychologists; some psychologists assume a feature set constructed or defined by the experimenter (Medin & Wattenmaker, 1987), some shift the burden to the subjects by asking them to list features (Rosch & Mervin, 1975), and some have little to say about the composition of the feature set (Tversky, 1977). Indeed, people in a new learning task may face a similar problem; of all the representable features, what is germane to encode? Prior experience will provide some guidance, but this will be increasingly inadequate as the task is increasingly novel. Perhaps, people are initially rather indiscriminate in the features encoded, including uninteresting features as well as features useful in either discriminating individuals or capturing regularities. Focused sampling directs attention away from the unsystematic, uninteresting features, that is, those which are not predictive at either the individual or group level. Thus, for people in novel learning situations focused sampling might be extremely useful for guiding selective encoding, by separating those features which might plausibly be useful from those which are useless.

Focused sampling is clearly not the only process involved in concept and rule learning. For example, additional operations are also needed so the learner is not irrevocably locked into examination of the same intensively
learned features. Although focused sampling was introduced for its relevance to concept learning, methods need to be added for designating change in representation for new categories (Davis, 1985). Finally, focused sampling is largely data-driven and makes little use of prior knowledge or theory. However, the sets of rules which focused sampling derives from the data are likely to be particularly compatible with and useful for more theoretical analyses. Correlation and interpredictivity among features is a very good basis for inferring stronger domain-linked relations, for example, causality for theories about physical objects, intentionality for theories about people, and scope of control or government for linguistic theories.

From a psychological perspective we consider evidence that bears on actual use of focused sampling by people. Evidence for or against focused sampling is minimal, but there are several relevant findings which are consistent with it, and there is one set of direct tests. Most of these findings come from research using tasks with feedback, not the primary task domain of focused sampling. Before we ask whether people use focused sampling, we should ask whether people learn and use feature covariations at all. Although people's abilities to give magnitude estimates of correlations are not uniformly good, people are sensitive to co-occurrence of qualitative features. They show this when observing members of a single class (Medin, Altom, Edelson, & Freko, 1982) and in incidental learning (Lewicki, 1985), not just in tasks which require use of covariation, such as learning a biconditional rule in an explicit task (Bourne, 1970). Two findings on learning correlations with feedback are more related to focused sampling. People's sensitivity may be limited when there is little structure in input or in prior theory (Estes, 1986). Focused sampling would predict that discovery of correlational rules is most difficult when the rule occurs in isolation among a large set of unsystematic cues, as was the case in the Estes experiments. Zeaman, Campione, and Allen (1970) also found that the benefit from redundant relevant cues depends on the relation to irrelevant cues; for example, converting a previously irrelevant cue to be relevant consistently helped while adding a relevant but new cue did not necessarily benefit learning. Further, the idea of attentional learning is supported in tasks with feedback (Fisher & Zeaman, 1973; Lovejoy, 1966; Nosofsky, 1984; Zeaman & House, 1963).

The model's predictions about learning without feedback are novel because few models have been proposed for this domain. Again, the first issue is simply whether people can learn covariation rules without feedback or whether they are restricted to generalization via overall similarity; initial evidence for rule learning is available for simple categories of animals and people (Billman & Heit, 1986; Billman, Richards, & Heit, 1987).

Extending attentional learning models to include facilitation among predictors leads to the specific prediction of facilitation from clustered, interpredictive features for learning without feedback. This is the opposite of
findings in learning with feedback, where presence of a covarying feature interferes with learning the relation between target and feedback features. Evidence from observational learning of the syntactic categories of artificial grammars provides some more direct evidence for clustered feature facilitation in observational learning (Billman, 1983; Billman, Heit, & Dorfman, 1987). This series of studies directly tested for clustered feature facilitation by varying the structure available for learning syntactic categories. Subjects watched animated displays with descriptive sentences in an artificial language and tried to learn the language. Conditions differed in the number of intercorrelated cues which provided the basis for the syntactic categories. For example, in some conditions different sentence roles (phrase structure rules) covaried with each other and with agreement rules. In a comparison condition all the structure in the first condition was present and, in addition, phonological form, semantics, and syntactic marker elements covaried. The learning of an identical rule present in both conditions was compared. Clustered feature facilitation predicts that the same rule will be learned faster in the system with additional rules among an overlapping set of features. A series of experiments found the predicted facilitation for learning the identical component when the component was part of the more complex grammar.

Finally, patterns in natural language acquisition are consistent with the benefit of multiple, interpredictive features, though this data is subject to many interpretations. The Hebrew gender system is characterized by multiple, consistent co-occurrence patterns among a small set of features; it differs from Germanic and Slavic systems in the regularity and transparency of marking. Hebrew gender is largely mastered very early (less than two years), much earlier than the less consistently marked systems (Levy, 1983). Slobin’s report of early use of the case marking system in Turkish (Slobin, 1982) is also consistent with the importance of multiple, regular, interpredictive cues to a system.

**GENERAL CONCLUSION**

Focused sampling is an adaptive plan or inductive bias which facilitates discovery of interrelated correlational rules. Discovery of interrelated correlational rules is of particular interest because they might be used to guide category formation, particularly in relatively implicit tasks where no feedback is available. Since natural inductive problems can be very complex, adaptive learning procedures are needed to prioritize which rules or regularities are explored first. The novel thrust of the present work lies in showing how attentional learning procedures can serve to constrain induction appropriately for learning categories with complex and partially correlated characteristics.
REFERENCES


