Journal of Experimental Psychology: General
1995, Vol. 124, No. 2, 181-206

0096-3445/95/$3.00

Causal Models and the Acquisition of Category Structure

Michael R. Waldmann
Max Planck Institute for Psychological Research

Keith J. Holyoak and Angela Fratianne
University of California, Los Angeles

This article proposes that learning of categories based on cause–effect relations is guided by causal models. In addition to incorporating domain-specific knowledge, causal models can be based on knowledge of such general structural properties as the direction of the causal arrow and the variability of causal variables. Five experiments tested the influence of common-cause models and common-effect models on the ease of learning linearly separable and nonlinearly separable categories. The results show that causal models guide the interpretation of otherwise identical learning inputs, and that learning difficulty is determined by the fit between the structural implications of the causal models and the structure of the learning domain. These influences of the general properties of causal models were obtained across several different content domains, including domains for which subjects lacked prior knowledge.

Tasks as apparently diverse as classical conditioning, category learning, and causal induction often require the learner to combine multiple cues in order to elicit a response. The cues may be conditioned stimuli (in conditioning), features of category instances (in category learning), or possible causes (in causal induction). Numerous learning models have been proposed in each of these areas, and a great deal of theoretical interest has focused on the extent to which common learning mechanisms may operate across these formally similar tasks. Most of these theories model learning as a domain-general process, bottom-up and basically associative in nature, that applies across diverse domains. Recently, more top-down or theory-based approaches have been proposed, which view learning as guided by domain-specific theories. In the present article we outline a position that is intermediate between these two views. We claim that a major subset of learning situations—those that the learner interprets as involving cause–effect relationships—are guided in part by domain-general knowledge, based on people's tacit understanding of causal relations. Learning proceeds by integrating top-down knowledge at different levels of generality with the bottom-up information provided by the environmental input to the causal induction process. Our view concurs with associationist theories in postulating sensitivity to domain-general properties of learning tasks. However, we also show that learning is guided by top-down knowledge about causal relations in ways that are not captured by current theories of associative learning. In this respect we agree with the focus of the domain-specificity view on the importance of top-down influences from prior knowledge.

Our testing ground is the ease of learning different types of category structures. We argue that the causal-model theory we propose can account for otherwise puzzling variations in the ease with which people learn categories on the basis of feature correlations that give rise to configural properties. We briefly review general associationistic models, as well as theories focusing on domain-specific knowledge, that have been proposed to account for learning of different category structures. We then describe our causal-model theory and report the results of five experiments that test its predictions.

Configural Learning

In multiple-cue contingency learning, the categories to be acquired are linearly separable (LS) if the optimal response can be generated as a linear function of the weighted cues (Minsky & Papert, 1969). For nonlinearly separable (NLS) categories, in contrast, the optimal response depends on relationships between cues, rather than on individual cues. The classic example of a nonlinearly separable task is the "exclusive-or" (XOR) problem (see Rumelhart, Hinton, & Williams, 1986). Here the system has to learn to generate

Michael R. Waldmann, Max Planck Institute for Psychological Research, Munich, Germany; Keith J. Holyoak and Angela Fratianne, Department of Psychology, University of California, Los Angeles.

Most of the research presented in this article was planned and conducted during a visit of Michael R. Waldmann to the University of California, Los Angeles, which was made possible by a grant from the Deutsche Forschungsgemeinschaft. This research was also supported by National Science Foundation Grant SBR-9310614. Portions of this research were presented at the 1989 meeting of the Psychonomic Society in Atlanta, Georgia; the 1990 Conference of the Cognitive Science Society in Cambridge, Massachusetts; the 1992 meeting of the European Society for Cognitive Psychology in Paris, France; and the 1993 meeting of the Psychonomic Society in Washington, DC. We thank Patricia Cheng and Eric Melz for helpful discussions.

Correspondence concerning this article should be addressed to Michael R. Waldmann, Max Planck Institute for Psychological Research, Leopoldstraße 24, 80802 Munich, Germany. Electronic mail may be sent via Internet to waldmann@mpipf-muenchen.mpg.de.

one response when one of two input cues is present (0,1 or 1,0) and another response when both are present or absent (1,1 or 0,0). There is no set of weights for two such cues that could make the system generate a positive response with each single cue but suppress such a response when both cues are jointly present. To learn an XOR problem, a system would have to be sensitive to the configural property of two cues being positively or negatively correlated within each of the two categories. An important limitation of simple two-layer connectionist networks is that they are unable to learn NLS category structures (Minsky & Papert, 1969).

The relative difficulty of LS versus NLS categories appears to depend on a variety of factors related to the task and learning material used. Research on animal conditioning has revealed the greater difficulty of NLS tasks than of corresponding LS tasks (e.g., Bellingham, Gillette-Bellingham, & Kehoe, 1985; Rescorla, 1972, 1973). Similarly, research on multiple-cue probability learning in which participants had to learn to combine several presented cues to generate a response has generally shown that participants have a harder time with configural predictors than with linear cue–outcome relations (see Brehmer, 1969; Edgell & Castellan, 1973; Mellers, 1980). By contrast, research on categorization presents a mixed picture. Whereas in some studies HNLS categories proved to be harder to learn than LS categories (Estes, 1986), other studies did not yield clear differences (Medin & Schwanenflugel, 1981). A classic study by Shepard, Hovland, and Jenkins (1961) found that an NLS category structure with two relevant features (Type II problem) was easier to learn than an LS structure with three relevant features (Type IV problem). Type II problems represent an XOR structure with an additional irrelevant feature. Type IV categories can be separated using a simple linear two-out-of-three rule: The presence of at least two of the three features indicates one of the two categories; otherwise the exemplar belongs to the other category. Shepard et al.'s results are particularly relevant in the present context because Experiments 1, 3, 4, and 5 use categories with Type II and Type IV structures. Difficulty of learning categories generally increases with the number of relevant features (Estes, 1986; Shepard et al., 1961). Linear separability is not the sole factor that influences ease of learning and may not be the major factor.

Associative Models

The inconsistencies in the empirical evidence concerning the relative difficulty of LS and NLS categories in human categorization have contributed to a proliferation of alternative learning models. Independent-cue models clearly cannot account for those results that show greater ease of learning NLS Type II problems than LS Type IV problems. Exemplar theories, which view classifications as being determined by similarity to stored exemplars, can account for Shepard et al.'s (1961) ordering of categorization difficulty when the models are augmented by parameters reflecting the degree of selective attention to the different category dimensions (Medin, 1975; Medin & Schaffer, 1978; Nosofsky, 1984, 1986).

Recently, a number of associationistic theories that model the actual process of acquiring categories have been proposed. Gluck and Bower (1988b) suggested that a two-layer connectionist network can provide a model of human categorization, with input units representing potential cues (such as symptoms of a disease observed in a patient) and output units representing classification responses (such as diagnoses of alternative diseases). The weights on associative links are learned incrementally using the least-mean-square learning rule (Widrow & Hoff, 1960). A similar model has also been proposed by Shanks and Dickinson (1987) for the learning of causal relations. They view causal learning as an associative learning process in which the cues represent potential causes and the responses are predictions of potential effects (see also Wasserman, 1990). The incrementally learned associative weights correspond to the perceived strength of the relationships between causes and effects.

Because a simple two-layer network is unable to learn NLS categories, Gluck and Bower (1988a) extended their adaptive network model. In their configural-cue model, they kept the simple least-mean-square learning rule but added explicit configural cues to the input layer, an assumption similar to the earlier proposal of Rescorla (1972, 1973). Gluck and Bower (1988a) showed that this model predicts that Type II problems are easier than Type IV problems, at least at the learning asymptote. One problem for configural-cue models is the potentially explosive growth of input nodes as the number of simple cues increases. Gluck, Bower, and Hee (1989) therefore suggested restricting the size to pairwise conjunctions of features. This restriction is empirically inadequate, however, because people can learn to classify items based on at least triples of features (Waldmann & Holyoak, 1990). Connectionist networks with hidden layers overcome the problem of exponential increase in network size, but at the cost of having to adopt more complex learning rules (e.g., back-propagation). Kruschke (1992) simulated learning with different types of back-propagation networks and found that NLS problems prove to be more difficult than LS categories. However, he could achieve a relatively good fit to Shepard et al.'s (1961) results by using an exemplar-based, connectionist learning model with additional weights that represent selective attention to the input dimensions. These weights are subject to learning, as are the other weights, and therefore are dependent on the structure of the stimulus–response mappings.

Although there are many important differences among the above models of learning (e.g., some stress abstraction of feature–category correlations, whereas others stress memory for specific instances), they are all domain-general, bottom-up, and associative in nature. Selective attention, which is assumed to be governed by the stimulus–response mappings in a bottom-up fashion, or by varying salience of cues, has typically been held responsible for the diverging empirical findings concerning configural learning. Unlike bottom-up models that focus on causal induction per se (e.g., Cheng & Novick, 1992; Kelley, 1967), none of the associative learning models is sensitive to the structurally distinct roles of causes and effects. The associative models

treat causal induction as simply a special case of contingency learning in general and therefore do not predict any influence of causal interpretations on the learning process.

Theory-Based Models

Associative models, which treat learning as a bottom-up process moderated by selective attention mechanisms, can be contrasted with a more mentalistic approach to learning, which claims that world knowledge guides the induction process. This view, which has a long history, has recently been resurrected in an influential article by Murphy and Medin (1985). Murphy and Medin contrasted *similarity-based* categorization theories, which view concepts as mere collections of features, with the *theory-based* view that concepts embody knowledge in which individual features are interconnected within a rich relational structure, partly based on unobservable causal factors. Research by Malt and Smith (1984) revealed that natural categories are distinguished not only by correlations between individual features and category membership (e.g., the property "can fly" is more likely to be true of birds than of mammals) but also by salient within-category correlations between features (e.g., among birds, those that live near the ocean are most likely to eat fish; see also Medin & Shoben, 1988). Murphy and Medin argued that such within-category co-occurrence relations between features may be based on specific, explicit causal knowledge (e.g., that being close to water is a precondition for catching fish; see also Murphy & Wisniewski, 1989).

Wattenmaker, Dewey, Murphy, and Medin (1986) investigated acquisition of different real-world categories to test how prior knowledge affects the difficulty of LS and NLS categories. In one set of experiments, they found that providing participants with a theme that encouraged the additive integration of features greatly facilitated learning of LS categories. For example, giving participants the hint that a category is related to the question of whether an object is suitable for use as a hammer encouraged the summing up of relevant features such as "easy to grasp" and "made of metal," which without the hint seem rather unrelated. As a consequence, a group that received the hint learned the task much more readily than did a control group without the hint.

In addition to supporting additive integration of features, prior knowledge may also support the use of interproperty relationships to predict category membership. In another experiment, Wattenmaker et al. (1986) compared the NLS Type II structure with the LS Type IV task. Participants had to learn to decide whether various persons were housepainters or construction workers. Without any further hint, the LS task proved easier than the NLS task. In the LS arrangement the participants simply had to add up features that were characteristic either of painters or of construction workers. However, the relative difficulty was reversed when participants received a hint that painters may be interior or exterior housepainters. This hint made participants sensitive to the correlation of the features "works inside" and "works year round" (interior housepainters) and "works outside" and "doesn't work in winter" (exterior housepainters). With this hint, the NLS Type II arrangement, which embodied this correlation, seemed more natural. (Also see Wattenmaker, in press.)

Other studies have also shown that the ease or difficulty of learning categories with various structural properties can be influenced by contexts that evoke prior explicit causal knowledge (Medin, Altom, Edelson, & Freko, 1982; Medin, Wattenmaker, & Hampson, 1987; Nakamura, 1985; Pazzani, 1991). In general, those studies yielded two major results: (a) People use prior knowledge about relations between specific entities in learning situations in which these specific entities are involved. For example, people who know that stretching balloons makes it easier to inflate them use this knowledge when categorizing balloons with respect to their inflatability (Pazzani, 1991). (b) People are generally insensitive to feature correlations unless prior knowledge about direct causal connections between these features is available (Medin et al., 1987).

Such studies have demonstrated that people are able to map their prior specific world knowledge to a learning task in a way that makes feature configurations more salient. An associationist might argue, however, that in such cases learners do not start with random or zero weights, but rather use previously learned associative links that are transferred to provide initial settings of the associative weights. Even though the details of this transfer process are far from clear, this approach does not seem incompatible with an associationistic learning paradigm (see Choi, McDaniel, & Busemeyer, 1993). In the present study we show that there are other types of knowledge-based influences that would be much more difficult to accommodate within associationist models.

Causal-Model Theory of Learning

Causal-model theory assumes that people will preferentially learn cause-to-effect relations, rather than effect-to-cause, cause-to-cause, or effect-to-effect relations. Focusing on cause-to-effect relations will reduce the cognitive complexity that would be involved in learning a complete matrix of covariation among all factors while still enabling predictive and diagnostic inferences (cf. Pearl, 1988). Causal-model theory shares with other theory-based approaches to categorization the assumption that people bring to bear prior knowledge on the learning task. But unlike previous accounts of this process, causal-model theory postulates that people derive top-down expectations not only from knowledge about specific causal relations (e.g., prior knowledge that stretched balloons are easier to inflate) but also from more general structural characteristics of causal relations. In the present article our focus is on two general properties that can guide the construction and use of causal models in learning a category: the direction of the causal arrow and the variability of the causal variables.

The most fundamental of these properties is the direction of the causal arrow. People assume that causes precede their effects and often observe this causal order, as when the start

of a fire is seen to precede the smoke it produces. Consequently, the causal arrow points from cause to effect and not the other way around. Observation of a known cause can trigger a predictive inference to the expected effect. However, people are often confronted with effect information that requires a diagnostic inference to an initially unobserved cause, as when observing smoke triggers the inference that there must be a fire. A fundamental assumption of causal-model theory is that, regardless of the temporal order in which we receive causal information, the underlying mental representation of the situation honors the cause-to-effect direction. The preference of people to learn in the cause-to-effect direction has been demonstrated in a number of studies (e.g., Eddy, 1982; Tversky & Kahneman, 1980; Waldmann & Holyoak, 1992; see also Einhorn & Hogarth, 1986).

The fact that order of observation can be decoupled from temporal precedence within the causal model provides the basis for our experimental dissociations between associationistic accounts and causal-model theory. In associative theories, learning typically implies updating of weights from inputs to outcomes. All associative theories, regardless of whether they see associative learning as a low-level process (e.g., Gluck & Bower, 1988b) or as modification of higher order beliefs (Shanks & Dickinson, 1987), share the assumption that the input corresponds to information obtained prior to outcomes. In a causal model, however, multiple input cues may be interpreted either as possible causes of initially unobserved effects (a *common-effect* model) or else as possible effects of initially unobserved causes (a *common-cause* model).

In a common-cause situation multiple effects are produced by a common cause (see Figure 1A). For example, symptoms of a disease might result from a common cause, such as a virus. In common-effect structures (see Figure 1B) the causal directions are reversed. Here multiple causes jointly converge on a common effect. For example, a number of individually causally insufficient facial cues might jointly be sufficient to produce an emotional response in an observer. Causal-model theory claims that causal connections will always tend to be acquired in the cause-to-effect direction, regardless of whether the nominal inputs (the cues for the required response) are interpreted as causes or as effects.

Levels of Causal Knowledge

Presenting identical learning tasks in the context of different causal models provides the basis for testing the predictions of causal-model theory against associationistic accounts of learning. An additional goal of ours in the present studies is to provide evidence for the use of domain-general causal knowledge. Causal situations can be represented on different levels of abstraction. On the most specific level, causal relations can be viewed as associations between specific event types. We know, for example, that high blood pressure is correlated with heart disease. On this level, the activated knowledge is tied to specific events (e.g., increase in blood pressure and onset of heart disease) within a specific domain (cardiovascular diseases). On a more abstract level, the same knowledge can be represented as involving a case of a cause and an effect. Here knowledge about causal directionality comes into play. Unlike domain-related knowledge about relations between two specific event types, knowledge about causal directionality is domain-general. On this more abstract level, high blood pressure is simply a particular token of the general class of cause events, and heart disease is a concrete manifestation of effect events. This abstract type of knowledge does not involve specific events; rather, knowledge about structural properties of causal structures is activated. For example,

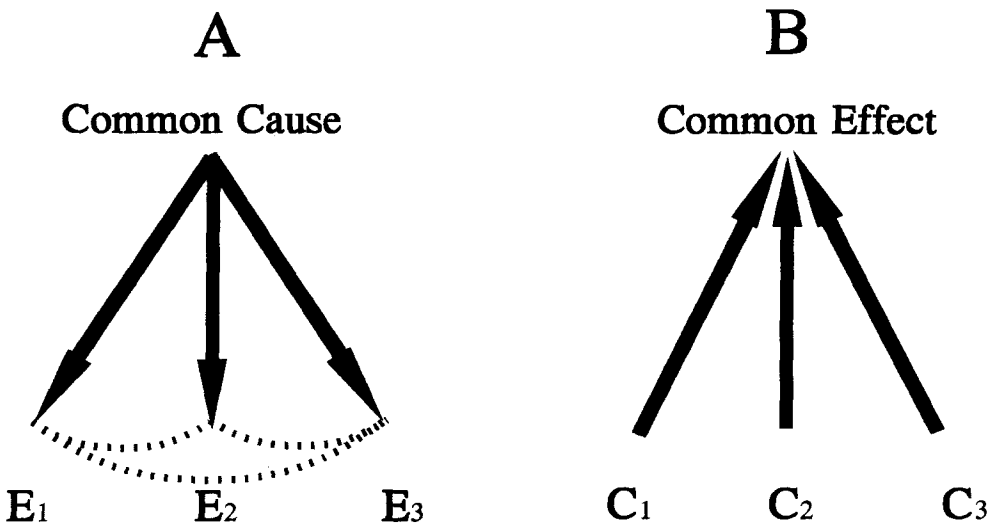

Figure 1. Common-cause structure (A) versus common-effect structure (B). Only the common-cause structure formally implies a spurious correlation (dotted curves) among effects.

regardless of how the two models in Figure 1 are instantiated, common-cause models imply spurious correlations among the effects, whereas common-effect models do not imply such correlations (as discussed in detail in the next section).

Previous research that has demonstrated the impact of prior knowledge on learning is ambiguous with respect to the level of causal knowledge that was actually used by participants. Typically, these studies investigated the effect of extra-experimental knowledge about specific correlated event types (e.g., changes in blood pressure and heart disease) on learning or categorizing (e.g., Medin et al., 1987; Pazzani, 1991). Although it may well be the case that the participants in these experiments activated abstract causal knowledge about causes and effects along with more specific knowledge about the particular events, the results of these studies are compatible with the more parsimonious hypothesis that learning was guided solely by specific knowledge. In the framework of associationist theories, the greater ease of learning about familiar events could simply be due to the fact that prior associative weights between specific events were transferred to a learning situation that involved those very same events. It does not appear necessary to invoke representations of abstract causal events or causal directionality in order to explain such empirical findings.

Given that domain-general causal properties are necessarily implemented within specific situations, how can domain-general knowledge experimentally be decoupled from more specific knowledge about causal relations? Our strategy in the following experiments was to present participants with situations for which they did not possess prior (extra-experimental) knowledge about causal relations. We then provided them with alternative instructions suggesting different patterns of causal directionality (common-cause versus common-effect models). For example, in Experiment 4 one group of participants read instructions indicating that magnets may affect the orientations of some surrounding iron compounds, whereas the instructions for a second group suggested that some of the iron compounds may causally affect the magnets. Previous work within the theory-based approach would predict sensitivity to correlations between magnets and compounds with both instructions. As in previous studies, this sensitivity may be due to knowledge about specific events that was acquired by means of the instructions, or it may be due to more general causal knowledge. Thus, sensitivity to the correlations among these particular events (e.g., magnets and compounds) is still ambiguous; it may be domain-specific. However, our goal was to show that participants are sensitive to additional aspects of the situations, which can be derived from causal models using knowledge about causal directionality and variability. Because in our experiments the causal models typically involved identical events (e.g., magnets and compounds), sensitivity to these additional aspects cannot be due simply to associative relations between these events; rather, such sensitivity would arise as a side effect of the representation of the direction of the causal arrow connecting these events and of the structure of the causal model.

Our goal is thus to demonstrate that people go beyond using direct causal knowledge about specific event types. We attempt to show that people also represent the direction of the causal arrow connecting these events, which is the basis for distinct patterns of interconnectivity among the events embedded in causal structures. In particular, we will show that participants are differentially sensitized to spurious correlations that have not been explicitly mentioned in the instructions but are nonetheless implied by some of the causal models. Because the participants had no prior extra-experimental experience with the causal situations described in the instructions, sensitivity to these additional, implicit aspects of the causal models cannot be attributed to participants' having had the opportunity to directly experience these properties. Rather, participants would have to derive such expectations from the more domain-general, structural characteristics of causal models. Our aim is to go beyond previous research (e.g., Medin et al., 1987) by demonstrating that causal models sensitize participants to feature correlations that are not supported by direct causal relations.

Our studies also differ from previous work in that we presented participants only with assumptions about potential causal relations. Participants had to use the learning input to test which of the suggested causal relations actually occurred; they could not simply transfer previous knowledge directly to the learning situation. The relation between knowledge and learning input is thus interactive (see also Wisniewski & Medin, 1994): Knowledge of potential causes guides the interpretation of the learning input, which in turn specifies or modifies the initial knowledge during the ongoing process of constructing a more accurate causal model.

Structural Implications of Common-Cause Versus Common-Effect Models

Reichenbach (1956) was one of the first philosophers to discuss the distinctions between different causal structures (see also Salmon, 1984). He postulated a *principle of the common cause,* which states that apparent coincidences too improbable to be attributed to chance can be explained by reference to a common antecedent cause. A famous example involves a group of actors who suddenly become ill after dining together. Even though there is a small chance that the actors contracted diseases independently of each other, it is more plausible to postulate food poisoning as a common cause. As Reichenbach pointed out, coincidences may be explained by a common cause, but not by a common effect.

In the present article we focus on situations in which causes and effects are psychologically continuous, that is, varying in intensity along some dimension. For example, it is easy to imagine a viral infection that can vary in intensity and trigger symptoms that also vary in intensity. People may represent such situations, in which a causal relation manifests itself in a correlation among quantitative variables, in terms of a common-cause model based on continuous variables. We assume people will bring to bear domain-general

knowledge that the strength of a cause typically covaries with the strength of its effects. With the formal apparatus of statistical causal-model theory, a common-cause model can be expressed in terms of structural equations (Bolles, 1989; Cartwright, 1989; Irzig & Meyer, 1987). For example, it may be the case that a virus C, which varies in intensity, independently affects the strengths of the symptoms E_1, E_2, and E_3 (as in Figure 1A). Thus, assuming standardized causal variables, when the virus is strong the three symptoms tend to be strong, and when it is weak the symptoms tend to be weak. The strengths of the three symptoms will clearly be correlated in spite of the fact that there are no direct causal connections between individual symptoms. The correlation is due solely to the influence of the common cause C. The following three equations formalize such a common-cause situation with three effect variables, E_1, E_2, and E_3, and the common cause variable C:

$$E_1 = w_1 \cdot C + U, \quad (1)$$

$$E_2 = w_2 \cdot C + U, \quad (2)$$

$$E_3 = w_3 \cdot C + U, \quad (3)$$

where w_1, w_2, and w_3 are causal weights expressing the strength of each causal relation, and U denotes an uncorrelated random error component. When the weights are nonzero, these three equations imply a spurious correlation between E_1, E_2, and E_3 as an implicit side effect of a causal structure that is fundamentally linear. Thus if response categories are defined by presence of the cause (in any degree) versus absence of the cause, and the cause takes on multiple values within the positive (i.e., cause-present) category, then a correlation among the effects within the positive category is predicted. An example of such a situation would be a learning task in which participants have to diagnose persons with a disease (Category A) against persons who did not contract the disease (Category B). The disease (Category A) is caused by a strong or weak virus. Accordingly, participants should expect to see patients with strong symptoms (E_1, E_2, and E_3 are strong) and they should expect to see patients with weak symptoms (E_1, E_2, and E_3 are weak) within the group with the disease (Category A). The expected strengths of the symptoms will therefore be correlated within Category A, which is defined by the presence of the disease. Of course, participants will have to use the learning input to infer which of the potential causal relations suggested in Equations 1–3 actually occur. We refer to the above representation as a common-cause model with a *varying* cause.

It is important to note that the common-cause model predicts a within-category correlation only in the case with a varying cause. A common-cause situation with a *constant* cause, in which the common cause takes on only one value within each response category, can also be represented in terms of Equations 1–3. The cause may be constant if it is discrete rather than continuous (the case discussed by Reichenbach, 1956), so that a causal factor is either present (Category A) or absent (Category B). Similarly, even if a causal factor takes on multiple levels that approximate continuous variation, only one level may be associated with each response category. For binary response categories the cause would be either at Level 1 (Category A) or Level 2 (Category B). An example of a common-cause model with a constant cause would involve a task in which participants have to classify patients into one of two disease categories, where Category A is produced by a strong form of a virus and Category B by its weak form. When the virus (C) is strong, Equations 1–3 yield a pattern in which the symptoms are strong, whereas when the virus is weak, the symptoms would be expected to be weak. It follows that in both types of common-cause models the case with the strong forms of the symptoms should be expected to occur within Category A. The case with the weak symptoms should be expected in Category A only when participants use a model with a varying cause; the same case should be expected in Category B when participants use a model with a constant cause.

In general, when the cause within each category is constant, no within-category correlation among the effects is implied by the model; however, each such effect will be correlated with a particular level of the causal factor (a *cue-to-category* correlation). Thus, common-cause models with constant causes should favor acquisition of category structures that embody such cue-to-category correlations.

The structural implications of a common-effect model, in which the causal direction is reversed as in Figure 1B, differ from those of the common-cause model with a varying cause. We will consider the symmetrical situation in which both cause and effect variables are continuous, but now the cues play the role of independent continuous causes converging in a linear fashion on an effect variable. If we have three cues, C_1, C_2, and C_3, producing the effect E, the situation can be described in Equation 4:

$$E = w_1 \cdot C_1 + w_2 \cdot C_2 + w_3 \cdot C_3 + U. \quad (4)$$

Whereas common-cause models with varying or constant causes imply different category structures, common-effect models with varying or constant effects both imply cue-to-category correlations, but not within-category correlations.[1] An example of a common-effect model with a varying effect would be a task in which one of the effect categories included both strong and weak intensities of magnetism and the cues represented continuously varying causes of magnetic strength. Similarly, a common-effect model with a constant effect would be constructed if one effect category was defined as high-intensity magnetism and the other as low-intensity magnetism. (Effects that must exceed a threshold to be observable will often approximate the case of a constant effect.) The crucial structural property of common-effect models is that they do not imply a within-category correlation between the causes, for the basic reason

[1] In some common-effect situations participants may have a tendency to average the influence of multiple causes (see Downing, Sternberg, & Ross, 1985). One can transform Equation 4 into an averaging model by dividing each weight by the sum of the weights. Note that common-effect models do not imply within-category correlations between the causes, regardless of whether participants are biased to add or to average the causal influences.

that effects do not influence their causes. Because in the simple network underlying Equation 4 the effect E does not itself cause anything, whether E is interpreted as constant or varying is irrelevant to any predictions regarding intercorrelations of the cues. Regardless of whether the three causes embodied in Equation 4 converge toward an effect varying in intensity within each response category or converge to produce multiple values of the effects across response categories, no interactions between the causes are implied by the model.

In a common-effect model, correlational relations among the causes would have to be formalized by explicitly postulating additional components that express types of interactions among the causes (e.g., $w_4 \cdot C_1 \cdot C_2$, as an example of a two-way interaction between the causes C_1 and C_2). Interactions among causes do, of course, occur in the real world. However, in such cases the underlying causal model would have to be elaborated with configural features to explicitly code the interactive relations between the causes and the effect. The need for explicit configural features would be expected to increase the difficulty of learning.[2] Thus in common-effect models, sensitivity to correlated causes will require explicit representations of interactive features; whereas in a common-cause model with a varying cause, such sensitivity can emerge implicitly from a causal network based solely on links from individual causes to individual effects.

How Causal Models Guide the Acquisition of Category Structures

One of our major goals in this article is to provide evidence that people are sensitive to the underlying causal structure of the learning domains. In particular, we are interested in whether participants make use of the implicit implications of causal models. We demonstrate this by showing that participants are differentially biased to learn different category structures when different causal interpretations are imposed on otherwise identical learning material. We vary category structure mainly by manipulating the property of linear separability in a particular way. Our choice of this manipulation does not imply, however, that the syntactic distinction between LS and NLS category structures provides a psychologically basic dichotomy between categorization tasks. Whether an LS or an NLS task is hard or easy to learn will, according to causal-model theory, depend on whether the particular category structure embodies a causal model that matches the assumptions made by learners. In principle, it is possible for a common-cause model with a varying cause to fit a category structure in which both within-category and cue-to-category correlations are present (a type of LS structure). However, presenting the within-category correlation in the context of NLS categories (as in Type II problems) has the methodological advantage that the within-category correlation is not redundant, but rather provides crucial information for learning the categories. Accordingly, in the present experiments we focus on differential learning of LS and NLS structures.

Two general assumptions of causal-model theory, which apply to situations in which participants have no relevant prior specific causal knowledge, are the basis for predicting the relative ease of learning different category structures: (a) Causal models structurally imply different causal situations. Thus, categories should be relatively easy to learn when they contain exemplars that match the expectations implied by the causal models. (b) The relation between causal models and the learning input is interactive. The learning input may lead participants to specify or modify their causal models. Categories requiring less change of the causal model that participants initially use to interpret the data should be easier to learn. Wisniewski and Medin (1994) recently demonstrated such tight couplings of theory and data in a different context. The above two assumptions imply that the fit between causal models and categories may vary in a continuous fashion. An assessment of the relative ease of learning two category structures therefore needs to be based on a comparison of their relative fit to the expected initial causal model. Even if both structures deviate from the predictions implied by the activated causal model, one may have a closer fit than the other.

The two category structures used in the present experiments may serve as an example of how the fit between category structures and causal models may be determined. One structure exhibited an LS cue-to-category correlation, the other an NLS within-category correlation. These two structures were special cases of Shepard et al.'s (1961) Types IV and II, respectively. Table 1 shows the assignments of eight exemplars, with Cases 1 to 4 corresponding to correct "yes" responses. Each case either has a high (H) or low (L) value on each of three dimensions. In Experiment 1, for example, these cases represent different persons, and the dimensions represent body signs of these persons that could vary in intensity (H or L). In the LS arrangement, high values of the three dimensions are more typical for the positive set, and low values are more typical for the negative set. For both sets, each dimension has one exceptional value, so that the dimensional values are only probabilistically related to the sets. However, a simple linear rule distinguishes the two sets: If a case has at least two out of three high values on the three dimensions, then the case belongs to the positive set. This linear rule leads to a clear-to-category correlation between the individual dimensions and the categories. High values of the dimensions tend to occur for positive instances ("yes" responses), whereas low values tend to occur for negative instances ("no" responses).

In the NLS condition, neither high nor low values are more or less typical for the positive or negative set. For each dimension, there are two cases with high values and two cases with low values in each set. There is no linear rule to

[2] Of course, these biases against interacting causes may be overridden by specific world knowledge when material from familiar domains is used. However, as pointed out by Dawes (1988), disordinal interactions are rare in our ecology, which may be why people do not expect them unless they have been prepared by specific knowledge.

Table 1
Structure of Item Sets in Experiments 1, 3, 4, and 5

Positive exemplars	Dimensions 1	2	3	Negative exemplars	Dimensions 1	2	3
Linearly separable							
Case 1	H	H	H	Case 5	L	L	H
Case 2	H	H	L	Case 6	L	H	L
Case 3	H	L	H	Case 7	H	L	L
Case 4	L	H	H	Case 8	L	L	L
Nonlinearly separable							
Case 1	H	H	H	Case 5	H	H	L
Case 2	H	L	H	Case 6	H	L	L
Case 3	L	H	L	Case 7	L	H	H
Case 4	L	L	L	Case 8	L	L	H

Note. Positive exemplars are correct "yes" responses; negative exemplars are correct "no" responses. H = high-intensity value on a dimension; L = low-intensity value on a dimension.

separate the two sets. In this category structure, the intensity of the dimensions is not correlated across the category boundaries: High and low values occur equally often on both sides. The only way to distinguish the two sets is to notice the positive correlation between the first and the third dimension within the positive set, which differs from the negative correlation in the negative set. The middle dimension is irrelevant for the classification. Thus, there is a perfect within-category correlation between the two relevant dimensions.

Let us now consider how each type of causal model would apply to the two category structures. It is possible to use Equations 1–4 to derive ordinal predictions about both the overall difficulty of the category structures and the relative difficulty of individual items. The item predictions are summarized in Table 2. The basic assumption is that item difficulty will be inversely related to the degree to which the relevant causal model predicts the correct response to an item. Each model generates clear predictions about the expected classification of the two extreme instances, HHH and LLL, which we term *prototypes.* The common-cause model with a varying cause (Equations 1–3) predicts that both HHH and LLL will be positive (i.e., cause-present) instances. In contrast, both the common-effect model (Equation 4) and the common-cause model with a constant cause (Equations 1–3) predict that HHH will be positive but LLL will be negative. We assume that to the degree an item is similar to a prototype and requires the response expected under the applicable causal model, it will be learned with few errors; conversely, to the degree that an item is similar to a prototype but requires the response opposite to that expected, it will yield more errors. (For the present category structures, these two influences are always in agreement with one another.)

As Table 2 indicates, this analysis yields from two to four clusters of items at different ordinal levels of predicted difficulty, depending jointly on the causal model and the category structure. The common-cause model with a varying cause influencing the positive items can be more readily fit to the NLS structure than to the LS structure, as discussed earlier. In Experiment 1, for example, participants in the common-cause condition were told that the dimensions represented potential effects influenced by a virus that could vary in intensity. Three clusters of items at distinct levels of difficulty are predicted. The HHH and LLL prototypes should be easiest because they are correctly expected to be positive. The HLH and LHL items each differ by one feature from a prototype and require the same response as predicted for the prototype (positive). The remaining four items also each differ by one feature from a prototype but require the opposite response (negative). The model suggested in the instructions to participants specifies only the positive set and hence is compatible with various kinds of negative exemplars, including one in which all the cases have normal values on each dimension (a situation that would correspond to an LS structure with a within-category correlation in the positive set). In our NLS structure, the negative exemplars are similar to the positive cases. In fact, before participants determine the irrelevancy of the second dimension, the negative cases and the positive cases HLH and LHL deviate equally (by one feature) from the two

Table 2
Predicted Difficulty of Individual Items in Each Category Structure as a Function of Causal Model (Experiments 1, 3, 4, and 5)

Model	Prototypes	Order of item difficulty (from least to most difficult): NLS category	LS category
Common-cause model with a varying cause	HHH and LLL are positive	{HHH_+, LLL_+} {HLH_+, LHL_+} {LHH_-, LLH_-, HLL_-, HHL_-}	HHH_+ {HLH_+, HHL_+, LHH_+} {LLH_-, LHL_-, HLL_-} LLL_-
Common-effect model	HHH is positive, LLL is negative	HHH_+ {HLH_+, LLH_-, HLL_-} {LHL_+, LHH_-, HHL_-} LLL_+	{HHH_+, LLL_-} {HLH_+, HHL_+, LHH_+, LLH_-, LHL_-, HLL_-}
Common-cause model with a constant cause	HHH is positive, LLL is negative	Same as for common-effect model	Same as for common-effect model

Note. Subscripts indicate items that are actually positive or negative. H = high-intensity value on a dimension; L = low-intensity value on a dimension. NLS = nonlinearly separable; LS = linearly separable.

positive prototypes. Accordingly, the NLS structure should initially be rather hard to grasp even with the help of the common-cause model. Because our instructions asked participants to identify the exemplars of the positive set (i.e., to diagnose the persons with the disease), we expected that participants would focus on the positive set. They should therefore try to fit the causal model to the positive cases, leaving the negative cases as the implicitly defined complement of the positive set.

The same model (common-cause model with a varying cause) is more difficult to fit to the LS structure, for which four levels of item difficulty are predicted. Prototype HHH is correctly predicted to be positive; items HLH, HHL, and LHH each differ by one feature from a prototype and require the expected response (positive); items LLH, LHL, and HLL differ by one feature from a prototype but are negative; and prototype LLL is unexpectedly negative. The initial causal model would have to be modified in order to be compatible with the learning input. One possibility would be to drop the assumption that the common cause is varying within the positive set, instead assuming that the three causal links express probabilistic relations between a constant cause and its three effects.

In the common-effect condition, the same dimensions were presented as in the common-cause condition. In this condition the three continuous dimensions were described as potential causes of a continuously varying common effect. In Experiment 1, for example, we told participants that the appearance of some persons may cause an emotional response in their observers. This description was designed to induce a common-effect model (Equation 4). Given that common-effect models based on individual causes do not imply interactions among causes, it follows that the NLS structure, which embodies a disordinal interaction, should be particularly hard to learn: The most dissimilar persons each cause the emotional response, whereas the intermediate persons do not cause it at all (i.e., the model predicts that the LLL prototype should be negative but it is positive instead). Four levels of item difficulty are predicted. Prototype HHH is positive as expected; the next cluster includes an item that differs by one feature from HHH and is positive (HLH), plus two items that differ by one feature from LLL and are negative (LLH and HLL); the next includes an item that differs by one feature from LLL and is positive (LHL), plus two items that differ by one feature from HHH and are negative (LHH and HHL); finally, prototype LLL is unexpectedly positive.

For the LS structure, the common-effect model provides a much more satisfactory fit. The model predicts just two difficulty levels: HHH is positive as expected and LLL is negative as expected; all of the remaining six items either differ by one feature from HHH and are positive or else differ by one feature from LLL and are negative.

The common-cause model with a constant cause predicts the same ordering of relative item difficulty as does the common-effect model. If a common-cause model with a constant cause is assumed from the outset, the model would span both the positive and the negative category. Participants should start with the expectation that the HHH case is a member of the positive set, produced by a cause of high intensity, whereas the LLL case should belong to the negative set, within which the cause is expected to be at its low intensity. The latter expectation is better matched by the LS than the NLS structure. Fitting the model to the NLS structure would require major structural changes, such as formation of explicit configural features. It follows that the NLS structure should be harder to learn than the LS structure if participants initially apply a common-cause model with a constant cause.

In Experiments 4 and 5, we used instructions that suggested a common-cause model with either a varying or a constant cause. The instructions and materials of Experiments 1 and 3 were relatively ambiguous between the two variants. In such situations, we expected participants to use the learning input as the basis for selecting the appropriate variant of the general class of causal models suggested by the instructions.

We now report the five experiments in which we tested specific predictions derived from causal-model theory. All of the experiments used continuous variables. In the first four experiments we focused primarily on comparisons between contexts that should evoke common-cause versus common-effect models with varying causal variables, because these are the causal models that yield the most distinguishable predictions regarding the relative ease of learning LS versus NLS structures. In Experiment 5 we examined the case of common-cause models with constant causes.

Experiment 1

In Experiment 1 we used a multiple-cue learning task in which participants received descriptions of fictitious persons. In the common-cause situation, the features used for the descriptions were characterized as potential effects (i.e., symptoms) of a disease caused by a virus. Both the symptoms and the virus could vary continuously in intensity. In the common-effect context, the same features, also introduced as varying in intensity, were redefined as potential causes of an emotional response in observers of the described person. Both the virus and the emotional response were characterized as potentially continuous. Because participants in both conditions saw identical stimuli and had to learn isomorphic responses, associative learning theories would predict identical learning rates.

The manipulation of participants' causal models was crossed with two types of category structures. These two structures correspond to the LS and NLS category structures displayed in Table 1. The dimensional body signs that described the appearance of the persons could be in one of two states, high in intensity (H) or low in intensity (L). The four cases displayed on the left side of Table 1 corresponded to the positive set, that is, persons with the disease or persons who elicited the emotional response in their observers.

In the common-cause condition, participants were told that the three dimensions represented effects (i.e., symp-

toms), the intensities of which were potentially affected by a common cause, also varying in intensity, such as a viral infection. Depending on whether the common-cause context leads to an interpretation of the likely underlying causal structure in terms of a common-cause model with a varying or constant cause, participants could prove sensitive to either the within-category correlation embodied in the NLS structure or to the cue-to-category correlation embodied in the LS structure. One possibility was that participants might be led to think that only the positive set describes patients who contracted the disease, whereas the negative set describes persons with no disease or diseases caused by other viruses. Because the main task in the experiment was to diagnose the new virus, participants would be expected to focus on the positive set. Furthermore, because they were told that the virus varies in intensity, they might expect a stronger and a milder form of the disease within this positive set. This interpretation leads to a common-cause model with a varying cause, in which a single continuously varying cause generates the positive instances and the negative instances are defined only indirectly as cases in which the cause is absent. These expectations are better matched by the NLS structure than by the LS structure, which would give the former structure an advantage during learning.

Alternatively, participants might view the positive set ("yes") as describing persons with a strong form of the virus and the negative set ("no") as persons with a weaker form of the virus. According to this model, all persons are affected by the virus, but the patients with the disease were probably exposed to its stronger form. This interpretation leads to a common-cause model with a constant cause, in which the two intensity levels of the cause are interpreted as a binary distinction between a level that probabilistically yields positive instances and one that yields negative instances. Participants who use such a common-cause model with a constant cause should be prepared to learn a category structure very similar to that embodied in the LS condition.[3]

The two variants of a common-cause structure thus respectively imply a within-category correlation within the positive set (as in the NLS structure) or a correlation between individual features and the optimal response (as in the LS structure). If the instructions are relatively neutral between the two interpretations, people may use the input to refine their basic common-cause model so as to best fit the observations. For example, if participants are told that case LLL requires a "yes" response (NLS condition), the common-cause model with a varying cause would seem more appropriate; whereas if they are told that LLL requires a "no" response (LS condition), the variant with a constant cause would seem more plausible. Assuming that people can use the input to adapt their common-cause model so as to best match the structure that is actually presented, it follows that participants in the common-cause condition should be able to learn either structure relatively effectively.

In the common-effect condition, the same features were presented as in the common-cause conditions. However, these features were introduced as potential causes of a *continuously* varying common effect, the emotional response of an observer. As pointed out in the Introduction, it was expected that common-effect models would not imply interactions among their causes. Thus, the NLS structure should be hard to learn relative to the LS structure, which corresponds to a linear causal situation with three independent, probabilistic causes.

Method

Participants. The participants were 40 undergraduates from the University of California, Los Angeles. Half of the participants were assigned to the common-cause condition and half to the common-effect condition. Half of each of these groups received the LS structure and half received the NLS structure. Assignment of participants to conditions was random.

Material. The stimuli were descriptions of people belonging to one of the two categories. The description consisted of three binary dimensional features: weight, pallor, and perspiration. On the basis of informal interviews, we tried to choose dimensions for which participants did not have prior knowledge about specific intercorrelations, which could sensitize them to correlations in the learning material. Each stimulus person had either high or low intensity values on each of these dimensions. The high values were "anorexic body," "ghostly white skin," and "sweating on face"; the corresponding lower values were "underweight," "pale skin," and "perspiring on forehead." Thus, unlike earlier studies (e.g., Medin et al., 1982) that used bipolar dimensions, we used features that were located on the same side of a neutral point. We selected such features in order to make the presented relation between the virus and each individual symptom more plausible. It seems more natural to assume that a virus varying in intensity will affect a dimensional symptom in a single direction, rather than cause opposite symptoms (e.g., being either underweight or overweight).

The eight possible cases were arranged in either an LS structure or an NLS structure. Table 1 shows the assignments of the eight cases, with Cases 1 through 4 corresponding to correct "yes" responses. Dimensions 1, 2, and 3 correspond to weight, pallor, and perspiration, respectively. Sixteen sets of these eight cases were prepared. Every description was typed on an index card. Each index card contained "Name" as a header followed by individual initials for each described person. By using different initials for each card, we led participants to believe that they were presented with a large sample of individual cases, rather than a small number of repeated cases.

Procedure. Participants were run in individual sessions. The material was arranged in a pile containing 16 blocks of the eight cases. The block structure of the material was not transparent to the participants. Within each block the cases were randomly ordered. Participants received index cards one at a time until they said "yes" or "no." After each response the experimenter gave "correct" or "incorrect" as feedback. Each description remained displayed for about 2 s after corrective feedback was given. Training continued until no errors were made on two consecutive blocks or until 16 blocks were completed.

[3] Because of the deterministic nature of the category/pattern assignments, the LS structure does deviate slightly from the normative output of a probabilistic common-cause model. In a truly probabilistic case, the negative cases (e.g., LLL) should in some rare instances also be produced by the strong form of the virus. However, we are only interested in the relative fit between causal models and the two tested category structures. Of course, there are other category structures selectively compatible with either of the causal models.

Participants in the common-cause condition were told that they were going to learn about a new disease caused by a new type of virus. They were told that the virus cannot be observed directly but that scientists are sure that it is a cause of observable body signs. Then the participants were told they were going to see descriptions of different patients and that half of the patients had contracted the new disease. We pointed out that not every body sign included in the descriptions was necessarily relevant for diagnosing the disease. Participants were instructed to try to get a general impression of the disease and its symptoms. Finally, we mentioned that the virus was not equally strong for all patients and that some patients were exposed to a very strong, very potent type of the virus so that they exhibited a relatively strong form of the disease, whereas other patients were only exposed to a weak type of the virus so that they were less strongly affected by the disease.

In the common-effect condition participants were told that in a series of psychological studies on interpersonal perception it had been found that people who observe other people sometimes react with a new emotional response to the physical appearance of the observed people. Participants were told that this emotional response is not directly observable but can be detected with a new psychophysical measuring instrument. Participants were then told that they were going to see descriptions of various body signs of different persons and that half of these persons trigger the new emotional response in observers. Participants were instructed to form a general impression of the causes of the emotional response, and we pointed out that not every mentioned body sign was necessarily relevant for predicting the emotional response. We also mentioned that the emotional responses could vary in intensity and thus that some people's appearances could trigger relatively strong responses, whereas other people could trigger relatively weak responses in the observers. Note that, despite this general hint about intensity variations, participants did not receive any feedback regarding the intensity level of the disease or the emotional response. Rather, participants were only given feedback concerning whether the outcome was obtained, regardless of its intensity.

Table 3
Mean Errors for the Eight Stimulus Types in Experiment 1 as a Function of Category Structure and Causal Context

Learning exemplars	Common cause	Common effect
	Linearly separable	
Positive items		
HHH	1.50	1.00
HLH	6.10	2.80
HHL	3.20	2.30
LHH	4.80	1.80
Negative items		
LLL	2.60	0.80
LLH	3.70	2.90
LHL	4.40	2.80
HLL	4.60	3.00
Average total errors	30.90	17.40
	Nonlinearly separable	
Positive items		
HHH	3.30	1.90
HLH	4.10	4.70
LHL	3.00	4.90
LLL	2.60	8.20
Negative items		
HHL	4.30	6.70
HLL	3.70	3.20
LHH	4.50	9.70
LLH	3.80	5.00
Average total errors	29.30	44.30

Note. H = high-intensity value on a dimension; L = low-intensity value on a dimension.

Results and Discussion

The mean total errors and the mean errors per item are shown in Table 3. A 2 (causal contexts) × 2 (structure of item sets) × 8 (items) × 16 (trial blocks) analysis of variance was computed, with decision error as the dependent measure. As predicted, the causal cover story interacted with the structure of the item set, $F(1, 36) = 7.55, p < .01$, $MSE = 2.10$. The LS and NLS structures were learned about equally easily in the disease context, which was expected to evoke a common-cause model, $t(36) < 1$. This equality is in accord with our assumption that bottom-up processing of the instances readily selects either the constant cause (LS) or varying cause (NLS) variant of the common-cause model. In contrast, in the emotional-response condition, which was expected to evoke a common-effect model, the LS set was easier to learn than the NLS set, $t(36) = 3.67, p < .01$. The advantage of the LS over the NLS structure for the common-effect model supports the assumption that people find linear main-effect models simpler to learn than models that code causal interactions. Overall, the LS set was learned with fewer errors than was the NLS set, $F(1, 36) = 5.95, p < .05$, $MSE = 2.10$.

Table 3 also displays the mean errors for the individual items. The relative difficulty of the eight item types varied significantly as a joint function of causal context and category structure, $F(7, 252) = 3.00, p < .01$, $MSE = 0.48$. For each of the four conditions, we used planned contrasts to test the significance of the predicted linear trend in mean errors across the item clusters derived from the theoretically relevant causal model, which were summarized in Table 2. For the common-cause condition, the constant-cause variant was the model predicted to be applicable to the LS structure. The mean numbers of errors per item for cases at the two predicted levels of item difficulty (ordered from easy to hard) were 2.05 and 4.47, respectively, $t(252) = 3.38, p < .01$. The varying-cause form of the common-cause model was predicted to be applicable to the NLS structure. The mean errors per item increased across the three predicted levels from 2.95 to 3.55 to 4.08, although the trend fell short of significance, $t(252) = 1.48, p < .20$. The item comparison between the LS and NLS structures for the common-cause conditions supports the hypothesis that the structure of the learning input triggers an appropriate causal model. Such bottom-up influences are most apparent in the pattern of performance for the critical LLL item. Learning that the LLL item belongs to the negative set in the LS condition proved just as easy as learning that it belongs to the positive set in the NLS condition (mean of 2.60 errors in each case). The structure of the learning input, and in particular the initial feedback for the LLL item, appears to have fostered generation of a common-cause model with either a constant

cause (LS condition) or a varying cause (NLS condition), based on general common-cause instructions.

For the common-effect condition, two ordinal levels were distinguished for the LS structure. These were ordered as predicted with mean errors per item of 0.90 and 2.60, respectively, $t(252) = 2.37$, $p < .05$. Finally, the common-effect condition predicted four levels of item difficulty for the NLS structure, which respectively yielded mean errors per item of 1.90, 4.30, 5.37, and 8.20, $t(252) = 5.27$, $p < .01$. Overall, then, the predicted ordering of item difficulty was observed in each of the four conditions.

Six participants in the LS common-cause condition versus 2 participants in the corresponding NLS condition, and 1 participant in the LS common-effect condition versus 5 participants in the corresponding NLS condition, did not reach the learning criterion within 16 blocks. As can be seen in Figure 2, the learning curves show systematic improvement across blocks of practice, $F(15, 540) = 22.9$, $p < .001$, $MSE = 0.17$. Inspection of Figure 2 reveals that superiority of the LS over the NLS structure is roughly constant throughout learning for the common-effect condition. For the common-cause condition, however, a crossover of the learning curves was observed. The NLS structure produced more errors than the LS structure on the first block, but after several blocks the relative difficulty of the two structures reversed. We statistically assessed the crossover by using orthogonal contrasts to evaluate the influence of causal context and stimulus structure for the first block of learning trials only. For Block 1, the NLS structure yielded more errors overall than did the LS structure, $t(540) = 3.25$, $p < .01$, with no interaction involving the common-cause and common-effect conditions, $t < 1$.

Why did the NLS condition start out as more difficult than the LS condition even within the common-cause context? One possible answer to this question makes use of our hypothesis that common-cause models can be generated for both category structures. A common implication of the common-cause models with constant or varying causes is that the HHH pattern should yield a positive response. As pointed out in the Introduction, participants may use the general heuristic with all causal models that the strength of the cause covaries with the strength of the effects. Using the HHH pattern as an initial prototype is more successful in the LS condition, in which at least two out of three dimensional values are high for each item within the positive set, than in the NLS condition, in which none of the three dimensions is individually predictive. In addition, even participants in the NLS common-cause condition who expect the HHH and the LLL cases to be affected with the disease should initially be at chance with the other cases because they do not yet know that one of the dimensions is irrelevant. The crossover suggests that the learning input guided construction of the appropriate causal model.

The crossover also rules out the possibility that the observed error pattern is due solely to differential initial biases concerning the association of individual items to correct responses, because the final pattern of item difficulty was not obtained in the initial learning block. The results there-

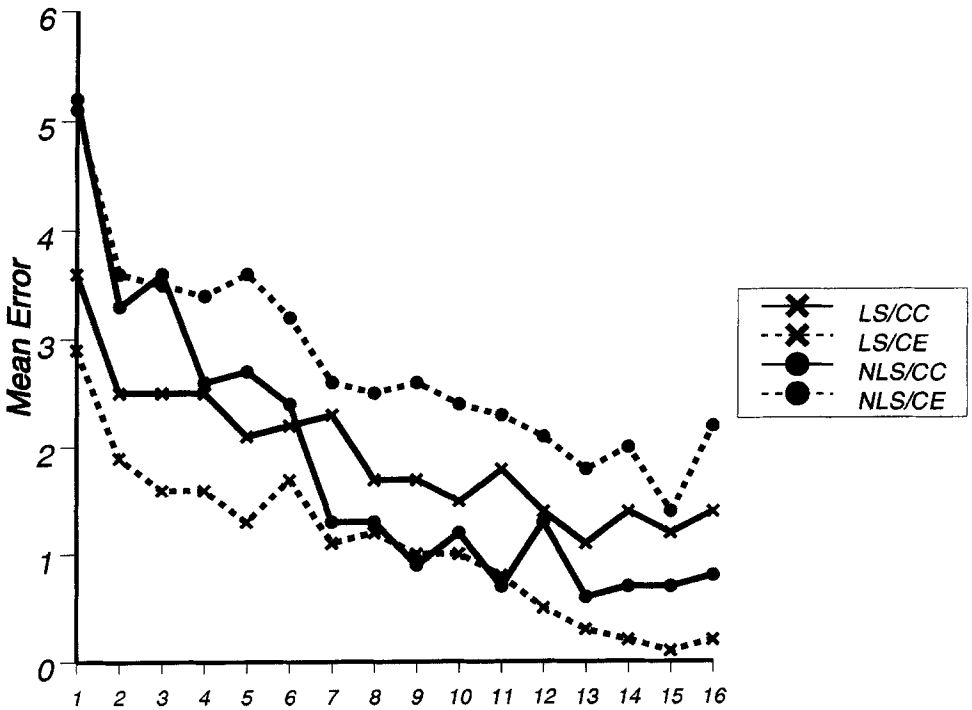

Figure 2. Learning curves in Experiment 1. Mean absolute error as a function of learning block, structure of item set, and causal context (maximum = 8 errors/block). LS = linearly separable; NLS = nonlinearly separable; CC = common cause; CE = common effect.

fore suggest that causal models guided the induction process beyond the initial learning block.

Experiment 2

In Experiment 2 we focused on an NLS structure. In order to better approximate correlations with continuous variables in testing our theory against associationist alternatives, we used category structures based on two variables with four intensity values each. A common-cause model with a varying cause predicts sensitivity to correlations regardless of the number of values within each dimension, because structures with corresponding relative ordinal positions of the values of each dimension produce the special case of a perfect positive correlation. In contrast, associative configural-cue models, in which each cue represents an arbitrary conjunction of feature values, do not capture the monotonicity involved in a correlation of continuous variables. Configural-cue models predict that there is no special advantage for a situation in which corresponding ordinal values define a category, as opposed to other more arbitrary configural structures. In addition, the number of configural cues required by such models grows as a power function of the number of levels. Causal-model theory predicts greater sensitivity to within-category correlations within a common-cause context (with a varying cause) than within a common-effect context, a prediction that cannot be motivated by associative learning theories even if they incorporate configural cues.

In addition to examining learning with more clearly continuous variables, in Experiment 2 we addressed the question of whether participants really need explicit information about the fact that the virus (the common cause) may vary in intensity. Even though capturing the positive correlation within the positive set requires the assumption of a continuously varying common cause, participants might be able to infer this property of the cause by observing the learning items. If the effects are clearly continuous (as was the case for our materials), this may encourage the assumption that the underlying cause is also continuous, rendering the overt hint unnecessary. Such a result would provide further support for the view that formation of causal models is partly guided by properties of the learning input.

Method

Participants. The participants were 40 undergraduates from the University of California, Los Angeles. Half of them were assigned to the common-cause condition and half to the common-effect condition. Half of each of these groups received a hint that the common cause might vary in intensity, whereas the other half did not. Assignment of participants to conditions was random.

Material and procedure. Stimuli were similar to the ones used in Experiment 1. The two dimensions of weight and pallor, with four intensity values each, were chosen for the descriptions of persons. The intensity variation again was restricted to one side of a neutral point. The four levels for weight were "slightly underweight" (Level 1), "underweight" (2), "seriously underweight" (3), and "anorexic body" (4). The levels for pallor were "mildly pale" (1), "moderately pale" (2), "very pale" (3), and "ghostly white" (4). Only an NLS structure was used. In the positive set ("yes"), which consisted of four different items, these two variables were perfectly positively correlated (i.e., values 4 4, 3 3, 2 2, 1 1). In the negative set ("no"), which also comprised four cases, the variables were negatively correlated (i.e., values 4 1, 3 2, 2 3, 1 4). These eight items were used for the learning task.

The procedure was similar to that used in Experiment 1. The disease instruction was again used for the common-cause condition, and the emotional-response instruction was used for the common-effect condition. The only difference was that in Experiment 2 it was pointed out to the participants that they were going to get information about two body signs, "the degree the person is pale, and the degree the person is underweight." For half of the participants the final hint that the virus or emotional response could come in different degrees of intensity was deleted from the instructions. The learning procedure was identical to that used previously, with the exception that participants received a maximum of 24 blocks of index cards.

Results and Discussion

The mean total errors and the mean errors per item are shown in Table 4. The manipulation of the causal context again proved effective. The NLS category structure was learned much more readily in the common-cause context than in the common-effect context. A 2 (causal contexts) × 2 (intensity hints) × 8 (items) × 24 (blocks) analysis of variance yielded a reliable effect of causal contexts,

Table 4
Mean Errors for the Eight Stimulus Types in Experiment 2 as a Function of Causal Context and Prior Information About Variability of the Intensity Level of Common Cause

Learning exemplars	Common cause	Common effect
	Without intensity hint	
Positive items		
Item 44	1.30	0.30
Item 33	2.20	4.30
Item 22	4.50	9.10
Item 11	3.90	8.90
Negative items		
Item 41	2.20	5.40
Item 32	3.50	7.10
Item 23	4.50	5.90
Item 14	3.90	4.40
Average total errors	26.00	45.40
	With intensity hint	
Positive items		
Item 44	0.90	0.10
Item 33	2.50	2.70
Item 22	2.50	10.30
Item 11	4.00	8.60
Negative items		
Item 41	1.60	3.80
Item 32	4.10	8.80
Item 23	3.10	6.90
Item 14	2.80	5.50
Average total errors	21.50	46.70

Note. In item labels, numbers correspond to intensity values of weight and pallor.

$F(1, 36) = 7.34$, $p < .025$, $MSE = 3.53$. Omitting the hint that the cause (virus) or the effect (emotional response) could vary in intensity did not significantly impair participants' performance ($F < 1$, for both main effect and interaction). The impact of causal models on learning an NLS structure thus generalizes to more continuous dimensions. Participants seem to be able to infer a continuously varying common cause on the basis of an observed correlation among continuous variables. This finding provides further support for the hypothesis that properties of the learning input influence the way the causal model is instantiated.

An alternative explanation of the results of Experiment 2 might be based on the assumption that participants generally assume variability of the causal variables within categories regardless of the properties of the input. One implication of this assumption would be that participants always infer a common-cause model with a varying cause, even when the variant with the constant cause is suggested in the initial instructions. In Experiment 5 we tested (and rejected) this prediction.

Table 4 also displays the mean errors for the individual items. The relative difficulty of the eight item types varied significantly as a function of causal context, $F(7, 252) = 6.18$, $p < .01$, $MSE = .34$. The specific item predictions derived in Table 2 for stimuli with binary-valued features do not apply to the stimuli used in Experiment 2, which have more continuous dimensions. In this experiment participants did not receive information about the number of levels of the causal variables and therefore had to induce their variability on the basis of the learning exemplars. Accordingly, predictions of the relative difficulty of items would be dependent on processes that go beyond the level of detail at which causal-model theory has yet been specified.

Nevertheless, the qualitative pattern of item difficulty was similar to that observed in Experiment 1. As in Experiment 1, the item with the strongest correlated features (here Item 44) was relatively easy to learn in both conditions. However, participants found it much easier to learn that the items with weak but correlated features (Items 22 and 11) were positive when the instructions established a common-cause rather than a common-effect context.

In the conditions without the intensity hint, 1 participant from the common-cause condition and 3 participants from the common-effect condition failed to reach the learning criterion within 24 blocks. The respective numbers for the conditions with the intensity hint were 1 for the common-cause context and 2 for the common-effect context.

Experiment 3

Medin et al. (1987) argued that people are sensitive to feature correlations only when they can bring to bear prior knowledge about specific causal links between these features. However, the results of the previous two experiments suggest that people can be sensitized to within-category correlations that are only implicitly coded with a common-cause model, even when they do not have prior knowledge of specific causal links. It remains unclear, however, how general these common-cause structures really are. We have proposed that causal models are based on structural properties of causal relations, such as causal directionality and monotonicity between causal strength and effect magnitude, that are relatively domain-general. So far, however, we have used only diseases as examples of common-cause structures. Although the results obtained in the common-effect conditions demonstrate that our participants did not have specific prior knowledge about direct correlations among the dimensions we chose in our experiments, it remains possible that participants made use of a domain-specific "disease schema." That is, it is possible that people generally know that symptoms tend to be correlated in the context of a disease, even when they do not have any further knowledge about a specific disease and its associated symptoms.

In order to bolster our hypothesis that people can represent novel learning situations in terms of causal models based on general structural properties, rather than just as instantiations of a more specific schema for diseases and their symptoms, in Experiment 3 we attempted to replicate the results of Experiment 1 using learning materials selected to be less familiar to participants. In this experiment participants learned about fictitious "moon stones." In the common-cause condition the stones were said either to have or not have a new substance called "zork" inside, whereas in the common-effect condition they were said either to elicit or not elicit pupil dilation in squirrels watching the stones.

The results of Experiment 1 suggested that when the instructions are relatively neutral between the variants of the basic common-cause model with constant or varying causes, participants can adopt whichever variant best fits the input, whether the input is an LS or an NLS structure. In this experiment the instructions specified more information suggestive of a varying cause, so that the NLS arrangement should have been easier to learn than the LS arrangement for participants in the common-cause condition. Participants were presented with continuously varying effect cues and were told that the common cause was present only within the positive set (the stones with zork). However, because the instructions did not mention that the amount of zork could vary within the positive set, some aspects of the instructions were also compatible with a common-cause model with a constant cause. We again expected that the learning input would help participants decide which variant of this causal model was more appropriate.

Method

Participants. Eighty undergraduates from the University of California, Los Angeles, participated in this experiment. The participants were randomly assigned to the four conditions, which resulted from crossing the causal context factor (common cause vs. common effect) and the structure of the item set factor (LS vs. NLS).

Material and procedure. Descriptions of fictitious moon stones were typed on index cards. As in Experiment 1, the description used three dimensions with two intensity values each. The dimensions used were color, size of spots on the surface, and

texture, which were typed in that order one below the other on the index cards. The high values were "dark blue," "large spots," and "very smooth"; the low values were "light blue," "small spots," and "slightly smooth." To obscure the fact that the pile of cards consisted of repetitions of eight types, we headed each index card by the label "Stone Identifier" with individual initials for each stone. The eight resulting item types were again arranged in either an LS or an NLS structure. These category structures were identical to those used in Experiment 1 (see Table 1).

In the common-cause condition, participants were told that American astronauts brought back a sample of stones from a recent excursion to the moon. Most of these stones were gray, had no spots on the surface, and had a rough texture. (We introduced the normal appearance of the stones in order to characterize the stones to be studied as abnormal, so that the critical dimension values would be interpreted as falling on one side of a neutral point, as did the features used in the previous experiments.) The instructions went on to say that in one area of the moon the astronauts found stones that looked different. Participants were told that scientists who studied these stones discovered that some of the stones had a core that contained a new type of substance, which they termed *zork* and were very interested in analyzing. Unfortunately, it was very difficult and expensive to find out which stones contained zork by breaking them, because they were extremely hard. However, researchers found that the zork affected the appearance of the stones containing it. The participants' task was to learn to judge whether or not the appearance of a stone was caused by the presence of zork.

The instructions in the common-effect condition were similar. But here the scientists were said to discover that the appearance of some of the stones had an interesting effect on squirrels: Whenever squirrels watched these stones, their pupils would dilate. In this condition, the participants' task was to learn to judge whether or not each stone caused dilation of squirrels' pupils. In both conditions, the three dimensions were mentioned in the instructions, which pointed out that not every feature included in the description of the stones was necessarily relevant to the classifications. As in Experiment 2, no hint or feedback was given regarding possible intensity variations of either zork or pupil dilation.

The learning task was identical to the ones used in the previous experiments. The upper limit of learning blocks in Experiment 3 was 20 blocks.

Results and Discussion

Table 5 displays the major results of Experiment 3. Most important, the NLS structure was again learned much more readily in the common-cause condition than in the common-effect condition. A 2 (causal contexts) × 2 (structure of item sets) × 8 (items) × 20 (blocks) analysis of variance with decision errors as the dependent variable yielded a reliable interaction between causal condition and category structure, $F(1, 76) = 8.82$, $p < .01$, $MSE = 2.78$. As in Experiment 1, within the common-effect condition the LS task proved easier than the NLS task, $t(76) = 2.08$, $p < .05$. Unlike the results of Experiment 1, the interaction obtained in Experiment 3 took the form of a clear crossover, because within the common-cause condition the NLS structure proved easier than the LS structure, $t(76) = 2.11$, $p < .05$. This advantage for the NLS common-cause structure is in accord with the fact that the instructions in Experiment 3, more clearly than those in Experiment 1, favored initial use of a common-cause model with a varying cause, which is more compatible with the NLS than the LS structure.

Table 5
Mean Errors for the Eight Stimulus Types in Experiment 3 as a Function of Category Structure and Causal Context

Learning exemplars	Common cause	Common effect
	Linearly separable	
Positive items		
HHH	2.95	1.80
HLH	4.10	3.65
HHL	5.05	3.45
LHH	3.65	4.20
Negative items		
LLL	2.25	2.40
LLH	5.25	4.05
LHL	4.30	3.95
HLL	5.45	4.50
Average total errors	32.95	28.05
	Nonlinearly separable	
Positive items		
HHH	1.90	3.40
HLH	2.55	5.50
LHL	2.65	4.75
LLL	1.60	5.50
Negative items		
HHL	2.50	5.60
HLL	2.75	6.25
LHH	2.60	5.80
LLH	2.20	5.10
Average total errors	18.80	41.90

Note. H = high-intensity value on a dimension; L = low-intensity value on a dimension.

Table 5 also displays the mean errors for the individual items. Overall, the pattern was quite similar to that observed in Experiment 1. For each of the four conditions, we used planned contrasts to test the significance of the predicted linear trend in mean errors across the item clusters derived from the theoretically relevant causal model, which were summarized in Table 2. For the common-cause condition, the varying-cause form of the common-cause model was predicted to be applicable to the NLS structure. The mean errors per item across the three predicted levels were 1.75, 2.60, and 2.50, respectively, a trend that fell short of significance, $t(532) = 1.51$, $p < .20$. In the case of the LS structure it is less clear which variant of the common-cause model would apply. The instructions were intended to encourage initial application of the varying-cause interpretation; however, given that some aspects of the instructions were more compatible with a constant-cause model, it is possible that at some point the stimulus structure would lead participants to shift to the more appropriate constant-cause variant. The main difference between the two variants at the level of individual items is that under a varying-cause model it should be especially difficult to classify the LLL item as negative, whereas this item should be relatively easily classified as negative under a constant-cause model. The data presented in Table 5 clearly favor the latter variant, because mean errors were low for both the HHH and LLL items

(2.58) and uniformly higher for the remaining items (4.63). As was the case for the comparable condition in Experiment 1, the trend test derived from the constant-cause variant was highly significant, $t(532) = 4.31$, $p < .001$, whereas that derived from the varying-cause version was not, $t < 1$. The item analyses thus suggest that even if the common-cause instructions initially weakly encouraged a varying-cause interpretation, participants succeeded in switching to the constant-cause version under the bottom-up influence of the LS structure.

For the common-effect condition, the four predicted levels of item difficulty for the NLS structure yielded mean errors of 3.40, 5.62, 5.38, and 5.50, respectively. Although only the HHH item was notably easier than the others, the overall trend was significant, $t(532) = 2.23$, $p < .05$. For the LS structure the two predicted levels yielded mean errors of 2.10 and 3.97, respectively, a highly reliable difference, $t(532) = 3.70$, $p < .01$.

Out of the 20 participants in each group, 6, 4, 3, and 1 did not reach the learning criterion in the LS common-cause, NLS common-effect, LS common-effect, and NLS common-cause conditions, respectively. Figure 3 displays the mean errors in each condition across the learning blocks. All groups improved their performance with practice, $F(19, 1444) = 81.0$, $p < .001$, $MSE = 0.15$, and the pattern of the learning curves was qualitatively similar to that observed in Experiment 1. Inspection of Figure 3 suggests that superiority of the LS over the NLS structure was roughly constant throughout learning for the common-effect condition. For the common-cause condition, however, the NLS condition produced more errors than the LS condition on the first block, but after two blocks a crossover of the learning curves was observed. As in Experiment 1, orthogonal contrasts performed only on data for Block 1 revealed that significantly more errors were made for the NLS than the LS structure, $t(1444) = 3.27$, $p < .01$, with no significant interaction involving causal context, $t < 1$.

Experiment 4

In Experiment 3 we replicated the results of Experiment 1 using more unfamiliar learning material, which provides further support for the view that participants are able to form causal models based on general structural properties when learning about new categories. However, even though we kept the learning cues constant in the different conditions, the experiments so far still do not completely rule out the possibility that participants brought some form of relatively general but nonetheless domain-specific knowledge to bear on the situation at hand. In the previous experiments, the two causal structures were always implemented in different content domains, so that it still can be argued that sensitivity to the different category structures was somehow attributable to prior assumptions about these different domains. It might be claimed, for example, that participants generally prove more sensitive to within-category correlations in the context of biological categories (e.g., diseases)

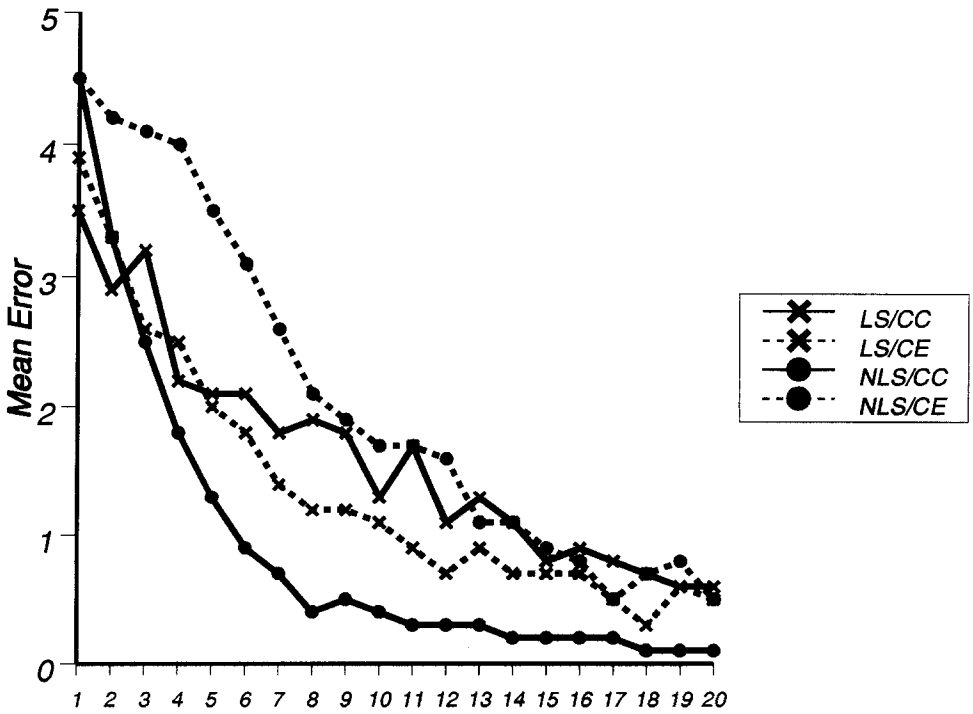

Figure 3. Learning curves in Experiment 3. Mean absolute error as a function of learning block, structure of item set, and causal context (maximum = 8 errors/block). LS = linearly separable; NLS = nonlinearly separable; CC = common cause; CE = common effect.

than in the context of social contexts (e.g., emotional responses; cf. Wattenmaker, in press). This hypothesis would be consistent with the pattern obtained in the first two experiments. Of course, such a claim is considerably weakened by the fact that essentially the same pattern of results was obtained in Experiment 3, in which the common-effect condition was realized within a biological domain. Nevertheless, stronger support for the causal-model theory could be obtained by comparing conditions in which different causal models are implemented within the same content domain. Ideally, all the entities used in the experiment should be identical except for the causal structure imposed on these entities. In that way possible content effects could be maximally controlled because the postulated causal models are manipulated by instructions. Waldmann and Holyoak (1992) presented an experiment (Experiment 3) that met these criteria but that did not manipulate category structure. In the present Experiment 4 we replicated the design of Experiments 1 and 3 using materials that controlled for content domain.

In contrast to the previous experiments, we used pictorial stimuli in Experiment 4. Participants were presented with pictures of fictitious stones brought back from Venus. Figure 4 shows one example of the learning stimuli, which displayed stones in the middle of dishes surrounded by colored iron compounds. Participants in all conditions received the same pictures and had to learn to judge whether or not the stone in the picture was a magnet by basing their decision on the orientations of the surrounding iron compounds.

In the common-cause context participants were told that scientists had discovered that some of these stones were either strong or weak magnets. In order to find out more about these stones, the scientists put the stones in dishes along with iron compounds. They found out that stones that were magnetic changed the orientation of some of the iron compounds placed in the dish. In the common-effect conditions, the same material was used but the direction of the causal connection between stones and compounds was reversed. Here participants were told that scientists had found out that some of the iron compounds emitted strong or weak magnetic waves that might magnetize the stones. The intensity of the magnetic waves was based on the orientation of the compounds.

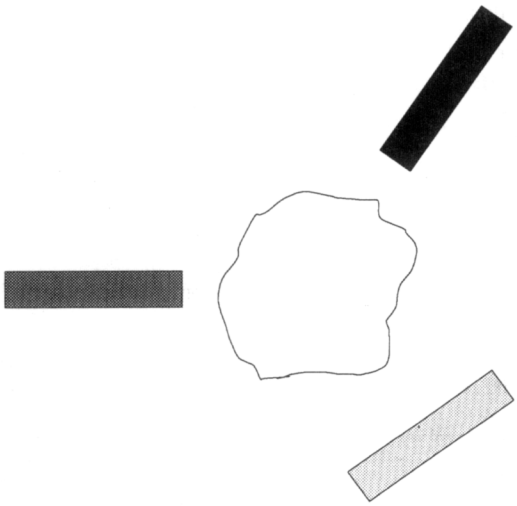

Figure 4. Example of the learning material from Experiments 4 and 5 (the HHL case): a potential magnet surrounded by iron compounds (compounds were blue, red, and green in the original set of learning items).

Even more explicitly than in Experiment 3, we attempted to provide instructions that would clearly favor a common-cause model with a varying rather than a constant cause, thus making the NLS arrangement easier to acquire than the LS arrangement within the common-cause condition. Participants were told that the common cause was present only within the positive set (the magnets) and varied in intensity within this set. They were also told how strong versus weak magnets affected the orientations of the surrounding compounds but that it was not necessarily the case that all of the compounds were affected by the stones. In addition it was pointed out that before the stones were placed in the dishes, the compounds were in a random orientation. (This implied that the orientation of compounds that were not affected by magnets was due to random factors.) We expected that these explicit instructions would favor a causal model for the common-cause condition in which a continuously varying cause was restricted to the positive set of stones. Thus, in the present experiment the instructions were designed to rule out a common-cause representation with a constant cause that would be compatible with the LS structure. More than in the previous experiments, these explicit instructions should diminish the potential role of the learning input in specifying the appropriate variant of the common-cause model. Experiments 1, 3, and 4 thus used instructions that increasingly specified aspects of the common-cause model with a varying cause.

The instructions for the common-effect condition were as parallel as possible. In the latter condition participants were told that there were strong and weak forms of the magnets (both to be judged "yes") and that there were stones that were not transformed into magnets ("no"). Participants were also instructed that not all of the compounds were necessarily causally effective and that the compounds were put in the dishes in a random orientation before the stones were placed in the middle.

Method

Participants and design. Forty undergraduates from the University of California, Los Angeles, participated in this experiment. They were randomly assigned to one of four conditions comparable to those used in Experiments 1 and 3.

Material and procedure. Participants received a series of index cards, each with a drawing of one stone placed on a dish in the middle of the card. Each stone was surrounded by three iron compound bars, represented by colored (green, red, or blue), narrow rectangular shapes. Each compound either had one of its ends pointed to the stone or had one of its sides facing the stone. Because different compounds could be identified by their colors,

they could occur in various positions on the index cards across learning trials. Also, even though the characterization of the orientations was binary, these orientations were defined relative to the stone in the middle so that the orientation relative to the index card could vary depending on the position of the compound. This display ensured that participants had to actively encode the meaning of the orientations of the compounds. (Pilot experiments in which we used dark or light colors in fixed positions had shown that with such material many participants seemed to resort to alternative visual learning strategies, such as memorizing the obviously restricted set of patterns.)

Sixteen sets of eight patterns were prepared for this experiment. As in Experiments 1 and 3, the LS arrangement of the eight patterns was compared with the NLS arrangement (see Table 1). Compounds with ends that pointed to the stones represented the high-intensity value, and compounds with sides facing the stones represented the low value. Figure 4 shows an example of an HHL case because two of the three compounds point to the stone. Within the NLS conditions, three counterbalancing groups varied which two of the three compounds (green, red, or blue) were correlated (with the 10th participant in each cell being randomly given one of these three assignments).

In the common-cause condition, participants received written instructions in which they were told that American astronauts brought back samples of stones from Venus. The instructions stated that back on earth, physicists discovered that some of these stones were magnetic in a way different from magnets found on earth. Participants were told that in order to find out which stones were magnetic, the physicists put individual stones in flat dishes along with different types of iron compounds. Also the instructions pointed out that before the stones were placed in the dishes, the compounds were in a random orientation. Believing that most types of iron compounds would be affected by the magnetic stones, the scientists tested dozens of different compounds on the stones. The research showed that stones that are magnetic are either strong or weak magnets. Strong magnets turn the magnetized compounds so that their ends point to the stone, weak magnets turn the magnetized compounds so that their sides face the stone. The instructions also stated that although most compounds are affected by magnets from Venus, not all compounds are affected. Finally, participants were told that they were going to see index cards, each with a different stone in the middle of the dish and the orientations of three randomly selected compounds represented by the colors green, red, and blue. Participants were asked to say "yes" when they thought the stone was a strong or weak magnet and "no" when they thought the stone was not a magnet. Before learning started, the task was summarized by the experimenter, who showed an example card with a stone and one brown compound. She explained how a strong or weak magnet would affect this compound and that the compound would simply stay in its random orientation if the stone was not a magnet or the compound was not sensitive to magnets from Venus. As a final test, participants were asked what the different orientations of the example compound could mean. They were encouraged to use this knowledge during learning.

The common-effect condition was closely modeled after the common-cause condition. The only difference was that participants in this condition were told that some of the stones from Venus could be transformed into strong or weak magnets. Research was said to show that the stones were magnetized by iron compounds. Iron compounds with their ends pointing to the stone send out strong magnetic waves, and iron compounds with their sides facing the stone send out weak magnetic waves to the stone. However, not all of the studied iron compounds turned out to be able to emit magnetic waves, and not all stones are affected by the compounds. Participants again were informed that they were going to see three arbitrarily selected compounds that had been placed on dishes in random orientations before the stone was placed in the middle. Their task was to say "yes" when they thought the stone in the middle of the dish had turned into a strong or weak magnet and "no" when the stone was not a magnet. Oral instructions and test questions were as similar as possible to those used in the common-cause condition. As in all previous experiments, participants did not receive any feedback regarding the intensity levels of either causes in the common-cause condition or effects in the common-effect condition.

The learning task was identical to those used in the previous experiments. The upper limit of learning blocks in Experiment 4 was 16 blocks.

Results and Discussion

Table 6 displays the results of this experiment. A 2 (causal contexts) × 2 (structure of item sets) × 8 (items) × 16 (blocks) analysis of variance with decision errors as the dependent variable again yielded a reliable crossover interaction between causal condition and category structure, $F(1, 36) = 11.4, p < .01, MSE = 2.82$. Whereas the LS structure was significantly easier to learn than the NLS structure within the common-effect condition, $t(36) = 3.22, p < .01$, the trend was in the opposite direction for the common-cause condition, $t(36) = 1.55, p < .20$. As in Experiment 3, we apparently succeeded in biasing participants toward ini-

Table 6
Mean Errors for the Eight Stimulus Types in Experiment 4 as a Function of Category Structure and Causal Context

Learning exemplars	Common cause	Common effect
	Linearly separable	
Positive items		
HHH	1.30	0.20
HLH	3.60	0.20
HHL	3.80	0.40
LHH	3.90	1.10
Negative items		
LLL	5.00	0.50
LLH	3.60	1.70
LHL	4.20	1.20
HLL	4.00	2.20
Average total errors	29.40	7.50
	Nonlinearly separable	
Positive items		
HHH	0.10	1.50
HLH	2.80	4.80
LHL	3.50	4.60
LLL	0.10	3.20
Negative items		
HHL	2.80	6.50
HLL	2.30	4.00
LHH	2.40	6.10
LLH	2.20	4.20
Average total errors	16.20	34.90

Note. H = high-intensity value on a dimension; L = low-intensity value on a dimension.

tially using a common-cause model with a varying cause, which fits the within-category correlation embodied in the NLS structure (by emphasizing that the cause, magnetism of the stone, only generated the positive and not the negative instances). The interaction between causal context and category structure observed in Experiment 4 cannot be explained by possible content biases, because the two contrasting causal models were implemented within the same domain. These results greatly weaken an alternative explanation of the results of previous experiments based on the hypothesis that participants rely on some kind of general content-bound knowledge of causal structures.

Table 6 also displays the error rates for the individual items. We again used planned contrasts to test the significance of the predicted linear trend in mean errors across the item clusters derived from the theoretically relevant causal model, as summarized in Table 2. For the common-cause condition, the varying-cause form of the common-cause model was predicted to be applicable to the NLS structure. The mean errors per item cluster across the three predicted levels were 0.1, 3.15, and 2.43, respectively, which yielded a significant linear trend, $t(252) = 3.70, p < .01$, by a trend test. The main difference in item difficulty was the greater ease of learning the two positive prototypes, HHH and LLL. For the LS condition, the item analysis provided additional evidence that the instructions used in Experiment 4 favored the varying-cause version of the common-cause model more heavily than the constant-cause version. The former variant predicts four levels of difficulty, which yielded mean errors of 1.30, 3.77, 3.93, and 5.0, respectively, $t(252) = 3.70, p < .01$, by a linear trend test. Most notably, the LLL item was especially difficult to classify as a negative case. Thus, unlike the comparable conditions of Experiments 1 and 3, the common-cause instructions used in Experiment 4 apparently not only biased participants to initially assume a varying-cause model but also led them to persist in applying this variant even in the LS condition in which the stimulus structure would be better fit by the constant-cause variant.

For the common-effect condition, the four predicted levels of item difficulty for the NLS structure yielded mean errors of 1.50, 4.33, 5.73, and 3.20, respectively. Although the error rate for the positive LLL item was lower than expected, the overall linear trend was significant, $t(252) = 2.51, p < .025$. For the LS structure the two predicted levels yielded mean errors of 0.35 and 1.13, respectively. This difference, although in the predicted direction, fell short of significance, $t(252) = 1.36, p < .20$. The item analyses thus again provided partial support for the predictions derived for the common-effect condition.

Out of the 10 participants in each group, 4, 2, 1, and 5 did not reach the learning criterion within 16 learning blocks in the LS common-cause, NLS common-cause, LS common-effect, and NLS common-effect conditions, respectively. Figure 5 depicts the mean error rates in each condition across blocks. The error rates for all groups decreased with practice, $F(1,540) = 15.25, p < .001, MSE = 0.13$. As in Experiments 1 and 3, the NLS structure yielded more errors overall in the first learning block than did the LS structure,

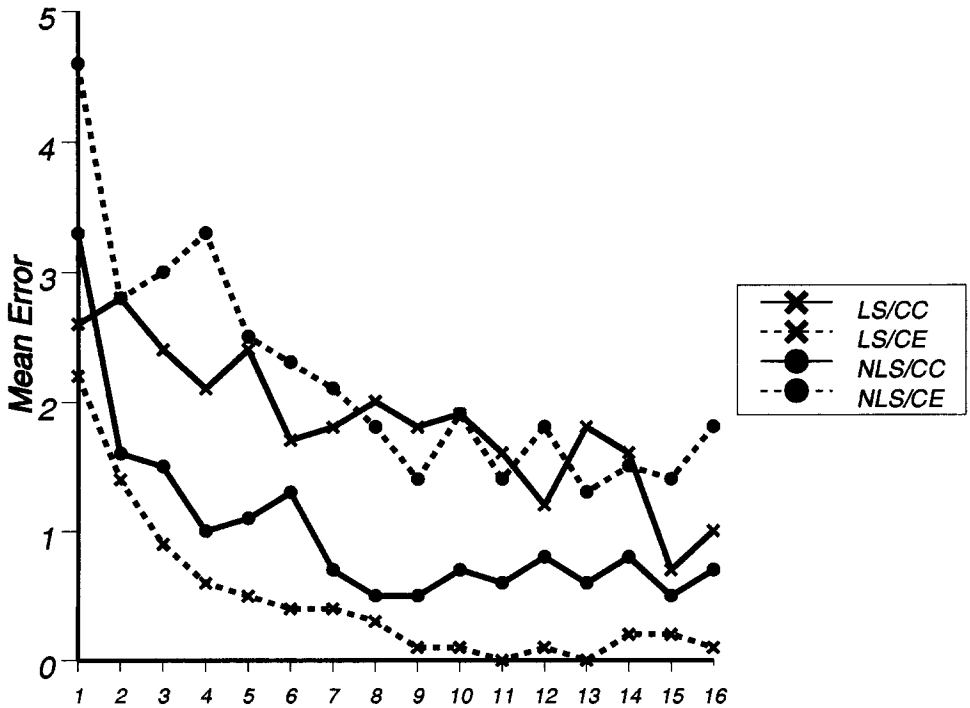

Figure 5. Learning curves in Experiment 4. Mean absolute error as a function of learning block, structure of item set, and causal context (maximum = 8 errors/block). LS = linearly separable; NLS = nonlinearly separable; CC = common cause; CE = common effect.

$t(540) = 2.40$, $p < .05$, with no significant interaction involving causal context, $t(540) = 1.32$, $p > .20$. By the second block the NLS common cause was easier than the LS common-cause condition, an advantage maintained across the rest of the learning blocks.

Experiment 5

We have argued that a common-cause model can be biased either toward a within-category correlation, as is embodied in the NLS structure shown in Table 1 (common-cause model with a varying cause), or to cue-to-category correlations between individual features and presence of a constant cause, as in the LS category structure (common-cause model with a constant cause). In previous experiments comparing these two conditions, the instructions were either neutral (Experiment 1) or designed to in some degree favor a common-cause model with a varying cause consistent with the NLS structure (Experiments 3 and 4). In Experiment 4 we used particularly explicit instructions to bias participants to form a common-cause model compatible with a within-category correlation. As we noted in the Introduction, we primarily focused on the common-cause model with a varying cause because it is this variant that yields a pattern of predictions (greater ease of learning of the NLS than the LS structure) that clearly differs from the predictions derived from a common-effect model. Experiment 5, however, was designed to test a further prediction of causal-model theory—that the relative difficulty of LS and NLS structures can indeed be reversed by manipulating expectations of the variability of the cause.

Accordingly, in Experiment 5 we modified the instructions so as to bias participants to generate a common-cause model with a constant cause, which should be compatible with the LS structure. We simply changed the mappings between the states of the continuous common cause and the two categories, so that the cause was set to a constant value within each category. The participants' task in Experiment 4 had been to say "yes" when they saw a strong or weak magnet and "no" when they saw stones that were not magnets. This task description should lead participants to expect both the HHH case (caused by the strong magnet) and the LLL case (caused by the weak magnet) to be in the positive set, an expectation that is better matched by the NLS than the LS condition. In Experiment 5, participants instead were instructed to say "yes" when they thought the stone was a strong magnet and "no" when they thought it was not a strong magnet, indicating that the two differing levels of the causal factor respectively generated positive and negative instances. This should lead participants to expect the HHH case and similar cases to be in the positive set (where the cause is strong) and the LLL case and similar cases to be in the negative set (where the cause is not strong). This expectation is better matched by the LS condition than the NLS condition, and therefore we expected the LS condition to be easier than the NLS condition in the present experiment. Thus whereas the common-cause instructions in Experiment 4 led participants to interpret the two intensity levels of magnets as approximations to a continuously varying cause operating on the positive instances, those in Experiment 5 led participants to interpret the intensity levels as a binary distinction between a level that produces positive instances and a level that produces negative instances. In other respects the instructions and the learning material were identical to those of Experiment 4. The common-cause model with a constant cause that participants were led to construct in Experiment 5 implies that strong magnets should cause compounds to be in their high-intensity orientation (e.g., HHH), whereas weak magnets should cause the affected compounds to be in their low-intensity orientation (e.g., LLL). This expectation matches the cue-to-category correlations embodied in the LS structure but is at odds with the NLS structure.

In Experiment 5 we also tested whether it was necessary to give participants prior information about the specific meanings of the orientations of compounds to obtain the predicted effects. In one condition, in which we adopted the procedure from Experiment 4 as closely as possible, participants were told which orientation signaled strong magnetic influences and which orientation signaled weak influences. In contrast, participants in a second condition were not given prior information about these assignments. We expected that these prior assignments would not prove necessary for participants to learn the LS condition, because the feedback should readily inform them about the meanings of the different orientations (because participants would assume that the orientations associated with "yes" responses corresponded to the higher intensity levels). However, providing information about assignments might be more important in the NLS condition, in which both orientations occurred equally often within both categories, so that the learning feedback did not provide participants with sufficient information to determine which orientation signaled a strong as opposed to a weak magnet. Of course, because the common-cause model with a constant cause that participants were encouraged to form would not match the NLS structure in any case, knowing the intensity levels associated with the orientations might still convey little or no benefit.

Method

Participants and design. Forty undergraduates from the University of California, Los Angeles, participated. All received common-cause instructions. Half of the participants were randomly assigned to the condition with LS categories and half to the condition with NLS categories. Half of each of these groups received prior information about the meanings of the orientations of compounds, and the other half did not receive such prior information.

Material and procedure. The material and procedure were virtually identical to those used in Experiment 4. The only difference was that in the present experiment participants were told that all stones were magnets and that they had to learn to say "yes" when they saw a strong magnet and "no" when they saw a magnet that was not strong. Again, participants were told that the compounds were randomly selected from a set that also contains compounds not affected by the magnets, and they were told that

before the stones were placed in the dish the compounds were in a random orientation. Participants who received prior information about the orientation of compounds were told that strong magnets turn magnetized compounds so that their ends point to the stone, whereas weak magnets turn magnetized compounds so that their sides face the stone. In contrast, participants who did not receive prior information were told only that strong magnets turn magnetized compounds in one particular direction and that weak magnets turn them in another particular direction. The assignments of the two orientations to strengths of magnets were counterbalanced in the conditions without prior information. Which two compounds were correlated in the NLS conditions was also counterbalanced (as in Experiment 4). The upper limit of learning blocks was again 16 blocks.

Results and Discussion

As can be seen in Table 7, the subtle changes in the instructions for Experiment 5, relative to the comparable conditions in Experiment 4, produced a massive switch in the relative difficulty of the LS and NLS conditions. A 2 (causal contexts) × 2 (intensity hints) × 8 (items) × 16 (blocks) analysis of variance revealed that the LS condition generated substantially lower error rates than the NLS condition, $F(1, 36) = 56.2$, $p < .001$, $MSE = 1.89$. This reversal of the difficulty of the two category structures was solely attributable to the different causal models participants were encouraged to bring to bear on the task in the two experiments, and not to differing learning experiences. Participants in both experiments received identical cues and identical feedback. (Recall that in Experiment 4 participants did not receive feedback about the strength of the magnets.) The results of Experiment 5 also serve to refute the possibility (discussed in connection with Experiment 2) that participants invariably assume a varying cause regardless of instructions.

Table 7
Mean Errors for the Eight Stimulus Types in Experiment 5 as a Function of Category Structure and Prior Information About the Assignments of Intensity Levels to Orientation of Compounds

Learning exemplars	With prior assignment	Without prior assignment
	Linearly separable	
Positive items		
HHH	0.00	0.40
HLH	0.50	1.20
HHL	0.30	0.90
LHH	0.10	0.40
Negative items		
LLL	0.00	1.10
LLH	0.90	1.00
LHL	0.70	1.20
HLL	0.10	1.20
Average total errors	2.60	7.40
	Nonlinearly separable	
Positive items		
HHH	1.00	4.70
HLH	4.40	5.40
LHL	6.40	6.00
LLL	5.90	3.30
Negative items		
HHL	7.40	5.30
HLL	6.00	4.30
LHH	7.30	5.70
LLH	5.40	5.30
Average total errors	43.80	40.00

Note. H = high-intensity value on a dimension; L = low-intensity value on a dimension.

No significant overall differences were obtained between the two information conditions; however, information condition interacted jointly with category structure and individual items, $F(7, 252) = 2.39$, $p < .025$, $MSE = 0.33$. Inspection of the error rates for individual items revealed uniformly low errors within the LS condition (see Table 7), so that cellar effects precluded finding any differences among classes of items. Regardless of whether prior information was given, all items could relatively soon be assigned to their correct categories. The NLS conditions yielded a different pattern. As can be seen in Table 7, the HHH item proved relatively easy in the condition with prior information about the assignment of intensity level to compound orientation, but not in the condition without prior information. This difference is in accord with the fact that responding "yes" to HHH patterns is consistent with the common-cause model with a constant cause that participants were biased to use. In the condition without prior information, the experimenter's feedback did not provide sufficient evidence to assign orientations to intensity levels, so participants could not possibly know which of the items represented the HHH pattern. Thus, this item could be uniquely identified in the LS but not in the NLS condition. For the condition with prior information, the common-cause model with a constant cause predicted four levels of item difficulty, which yielded mean errors of 1.0, 5.27, 7.03, and 5.9, respectively, $t(252) = 3.02$, $p < .01$, by a trend test. The much greater ease of classifying the HHH item (mean of 1.0 errors) versus the LLL item (5.9 errors), both of which are positive in the NLS structure, provides evidence that participants applied a constant-cause version of the common-cause model, rather than the varying-cause version (which predicts equal ease of classifying both prototypes).

Learning clearly improved with practice, $F(15, 540) = 13.09$, $p < .001$, $MSE = 0.14$. All participants in the LS condition learned the categories within 16 blocks, whereas 6 and 4 participants in the NLS condition with and without prior information, respectively, did not reach the learning criterion within 16 blocks.

General Discussion

Summary

The results of the present study indicate that providing people with very general structural information about potential causes and potential effects—information about causal directionality, continuity of causal variables, and the variability of the causal variables—allows them to construct causal models of the learning situation, which they then use

to induce the actual causal relations. A common-cause model with a continuously varying cause that generates only positive instances is compatible with a within-category correlation as embodied in an NLS structure; a common-cause model with a constant cause in which different levels of the cause generate positive and negative instances is compatible with cue-to-category correlations, as embodied in an LS structure. If the instructions are neutral between the two variants of the basic common-cause model, then the structure of the stimulus set appears to encourage generation of the appropriate variant. Using instructions that favored a common-cause model with a varying cause, we were able to show sensitivity to within-category correlations for continuous causal dimensions with more than two values. In two experiments, results indicated that participants can infer that a hidden cause was continuously varying on the basis of the observed instances of its effects, even without any explicit hint by the experimenter (see also Waldmann & Holyoak, 1990, Experiment 3).

In the first two experiments we used diseases as common-cause domains and emotional responses as common-effect domains. In Experiment 3, we were able to replicate the basic findings of the other experiments using less familiar learning material. In Experiment 4 we succeeded in replicating the same pattern yet again, using learning material in which the alternative causal models were implemented within a single content domain. These extensions support the view that people are able to utilize causal models, rather than only schemas that apply to more limited domains, such as diseases or biological entities.

In the present study we were able to show that the variants of the basic common-cause model with varying or constant causes favor acquisition of different types of category structures. An obvious question is whether such distinct variants could also be identified for common-effect models, which in the present experiments were invariably more compatible with the LS than the NLS arrangement. Could some form of common-effect model, in the absence of specific prior knowledge about causal links, favor an NLS structure? Within the range of simple one-step causal networks of the sort used in the present study, the answer appears to be no. As we noted in the Introduction, for a terminal effect in a causal network (i.e., one that does not causally influence some other factor being modeled), it is irrelevant whether the effect is constant or varying with respect to the response categories, because it has no causal impact. In this respect causal-model theory predicts an asymmetry between common-cause and common-effect models.

It is certainly possible, however, that some type of more specific prior knowledge could lead to generation of a common-effect model compatible with an NLS structure. For example, prior knowledge that encourages including explicit configural causes in a causal model (e.g., knowledge that drinking alcohol and taking a prescription drug may causally interact) could aid acquisition of an NLS structure within a common-effect model (see Footnote 2), as could prior knowledge that overrides the default assumption that cause–effect relations will be monotonic (e.g., knowledge that extremes on some dimension, such as body weight, may sometimes have similar causal consequences). It is also the case that interesting variants of common-effect models may emerge in the context of more complex causal situations involving causal chains and loops, in which the effects of initial causes can in turn have their own causal consequences. In general, much more work will be required to explore the range of causal models that people can construct on the basis of different types of information and the impact of different models on the ease of acquiring different types of category structures.

Causal Versus Noncausal Categories

Our main focus in this article has been on categories that lend themselves to causal representations. The presented experiments show that participants, even when they do not have specific world knowledge, may use more abstract types of knowledge that guide learning in a top-down fashion. Associative learning theories (including exemplar-based or prototype theories) are unable to explain these results because they generally focus on data-driven learning processes.

Nosofsky, Gluck, Palmeri, McKinley, and Glauthier (1994) recently presented a replication of Shepard et al.'s (1961) seminal study with artificial, geometric stimuli as learning material. Their results correspond to the results we obtained within the common-cause condition, with the LS Type IV problem turning out to be harder to learn than the NLS Type II problem. As these investigators show, this result is anticipated by a number of current categorization models including the configural-cue model (Gluck & Bower, 1988a), ALCOVE (Kruschke, 1992), and Anderson's (1991) rational model. One may speculate that this striking correspondence may be due to these models' being implicitly tailored toward natural kind concepts, which tend to embody common-cause models with a varying cause. For example, Anderson (1991) explicitly developed his theory of categorization with respect to "living objects."

Of course, learning is influenced not only by knowledge but also by properties of the learning input and by the learning strategies participants use. Many additional factors undoubtedly determine the relative difficulty of learning different category structures. Particular learning strategies, such as memory-based learning, may also sometimes predict sensitivity to category correlations (Wattenmaker, 1991, 1992, 1993). In addition, the characteristics of the cues may determine the semantic interpretation of the task. In the present experiments we mainly manipulated participants' causal models by providing different initial learning instructions, holding the cues constant across the different conditions. It is quite possible, however, that properties of category dimensions may influence the representation of the learning material in a way selectively compatible with different models. For example, although different causes that vary in intensity can easily be mapped to a common-effect model in which additively integrated intensity levels of the causes influence the effect, arbitrary values of artificial noncausal categories (e.g., square vs. circle) are less likely

to be plausible candidates for a model in which the features are integrated additively. Identical category structures may thus be represented differently depending on the characteristics of the features assigned to the categories. Within a memory-based or hypothesis-testing learning strategy, an LS category with three arbitrary features (a Type IV structure) may turn out to be harder to learn than an NLS category with two relevant features (a Type II structure), whereas the opposite may be true for features that can be integrated. More generally, the relative difficulty of different category structures can be expected to depend both on the semantic and pragmatic context of the task and on the learning strategies that people use.

Causal Models Versus Attention Weights

The results of the present experiments present problems for associative learning theories, because in all associative theories whatever information is first received is coded as cues on an input layer, which then triggers responses on an output layer. Accordingly, in predictive learning situations the cues always correspond to causes (e.g., Shanks & Dickinson, 1987), whereas in diagnostic learning situations the cues are mapped to effects (e.g., Gluck & Bower, 1988b; Shanks, 1991). This characteristic insensitivity to causal direction is the crucial feature of associative learning that the present results appear to refute. These models predict identical learning in situations with identical cues and outcomes regardless of whether the cues represent causes or effects. Insensitivity to causal direction is a property not only of simple associative learning theories but also of more complex connectionist theories (e.g., back-propagation networks) and concept theories of the similarity-based variant (exemplar or prototype theories).

One response to the present results, as well as to other recent evidence that causal directionality influences learning (Waldmann & Holyoak, 1992), is to develop network models in which the causal interpretation of events guides the assignment of information to layers of the network. That is, cues that are interpreted as "causes" could be assigned to the "input" layer, and cues that are interpreted as "effects" could be assigned to the "output" layer, regardless of the temporal order in which the cues are overtly presented (see van Hamme, Kao, & Wasserman, 1993). Links would then be interpreted as directed causal connections. For example, the causal networks proposed by Pearl (1988) and Peng and Reggia (1990) have this general character. Of course, the generation of responses then becomes much more complicated than in standard associative networks, because the responses could not simply be elicited by the input information, as in a standard feed-forward connectionist architecture. Rather, in diagnostic tasks the inputs would have to be mapped to the output level of the causal network. The activation pattern on the output layer would then have to be interpreted as being caused by an unseen input, which would have to be induced by diagnostic learning. Although it may well be possible to develop a network model of causal learning along these lines, it should be clear that some nontrivial problems must be solved to account for learning within diagnostic contexts. Such models would go well beyond simple associationism and in fact would instantiate the causal-model theory we are advocating.

Is there some way in which associative learning theories could accommodate our evidence of differences in learning rates for isomorphic learning structures without explicitly incorporating a causal semantics (which would render them equivalent to the causal-model approach)? As our earlier review indicated, inconsistent results concerning the relative difficulty of different stimulus structures have emerged from previous research (even though researchers have typically focused on one paradigm while ignoring contradictory results in others). For example, animal learning experiments and studies of human multiple-cue probability learning have typically shown a learning advantage for LS structures, whereas categorization research presents a mixed picture, with some studies even finding an advantage for NLS structures over LS ones. Within associative theories, variations in learning difficulty have typically been explained by fitting attention or learning-rate weights associated with configural cues to the results at hand. These models have treated such weights as being regulated by properties of the input—that is, they view the selection of weights as strictly governed by bottom-up influences. An obvious tack one could take, given results such as those presented here, is to introduce top-down influences on weight selection. In particular, one might postulate higher weights for configural cues within common-cause structures with a varying cause than within common-effect structures. (Of course, this idea is considerably weakened by the fact that common-cause models with constant causes would then require the assumption of relatively low weights for configural cues.) It is hard to see how such a move can be motivated without resorting to a formal analysis of the different structural implications of causal structures, which is outside the realm of current associationistic theories.

In a different set of studies, we have collected converging evidence for causal models that cannot be explained by postulating arbitrary changes in attention or learning-rate weights (Waldmann & Holyoak, 1992). In these experiments, we used a blocking paradigm. Blocking experiments typically make use of a two-phase learning design (e.g., Kamin, 1969). In Phase 1, participants learn to predict an outcome on the basis of a single valid cue. In Phase 2, a second redundant cue is constantly paired with the already established valid cue. All current theories of associative learning predict that participants should afterward be reluctant to predict the outcome when confronted with the second cue alone, even though it is perfectly correlated with the outcome within Phase 2 of the learning task. The Rescora–Wagner or least-mean-square learning rule, for example, predicts only an updating of weights when something unpredicted occurs. Because during Phase 1 participants learned to predict the event perfectly on the basis of the valid initial cue, no further learning should occur to the redundant cue in Phase 2. Note that this prediction holds regardless of how large or small the learning rates or attention weights for the cues are set. Thus even a "causally enriched" theory of associative learning, which postulates regulation of such weights by a front-end causal theory,

would predict blocking for common-cause as well as common-effect structures. However, Waldmann and Holyoak (1992) found cue competition only when the cues corresponded to causes (i.e., when instructions encouraged participants to form a common-effect model), not when they corresponded to effects (i.e., when instructions encouraged a common-cause model), which supports the prediction of causal-model theory (see also Melz, Cheng, Holyoak, & Waldmann, 1993). As in the present set of studies, these predictions of causal-model theory can be derived from the assumption that participants form causal models that are directed from causes to their effects. Part of the reason for the asymmetry of cue competition in the diagnostic and predictive tasks is the fact that multiple effects of a common cause do not interact, whereas there is a potential interaction between multiple causes of a common effect.

What Is General in Causal Models?

The theory-based view of categorization (Murphy & Medin, 1985) has emphasized the importance of domain-specific causal knowledge in guiding category induction (e.g., Medin et al., 1987). Prior knowledge of specific causal relations, when it is available, is clearly an important influence on subsequent learning. However, a complete model of causal induction must be able to specify how novel causal relations can be acquired even when specific prior knowledge is lacking. Causal-model theory, we believe, helps to close the gap between learning based on transfer of specific prior knowledge and learning based on bottom-up analysis of the current input. As noted earlier, one can potentially transfer specific prior knowledge to a new learning task by biasing the initial weights for particular cause–effect links within a causal model. However, the present study provides evidence that even in the absence of such specific prior knowledge, people can use more general structural information to form a causal model that implicitly biases the causal induction process in predictable ways. Participants in the common-cause condition of Experiment 4, for example, had no prior knowledge to allow them to predict which of the iron compounds would respond to the magnetism of stones from Venus. Nonetheless, the instructional context conveyed that (a) magnets had a causal influence on the orientation of some of the compounds, rather than the reverse, (b) the degree of magnetic influence could vary continuously, and (c) the orientations could vary even when no magnet was present. This information proved to be sufficient to allow participants to set up a common-cause model with a varying cause. Once this model was formed, sensitivity to the sort of within-category correlation embodied in the NLS stimulus structure naturally emerged. The causal mode provided a partially specified top-down framework within which bottom-up induction could operate for the acquisition of knowledge about specific causal connections.

The analyses of the learning curves obtained in Experiments 1, 3, and 4, which proved strikingly similar, suggest that causal models interacted with the learning input during the induction process. If participants had simply made use of prior knowledge about specific interproperty correlations, these biases presumably would already have been evident during the first learning block. Instead, the obtained learning curves revealed reliable crossovers in the difficulty of the LS and NLS structures for the common-cause condition, which suggests that the causal models provided only tentative hypotheses about the general form of causal relationships, leaving the specific causal relations to be induced through bottom-up analyses of the inputs. This result can be interpreted as an example of a "tight coupling" between the knowledge component and the learning component, similar to effects found by Wisniewski and Medin (1994). These investigators have shown that learning models that separate an associative learning component from a knowledge component are implausible, because these two components seem to interact during the entire induction process. In their experiments, Wisniewski and Medin demonstrated that the interpretation of category features is crucially dependent on the type of intuitive theory participants bring to bear on the task.

One of the most important computational advantages of models based on causal networks is that the statistical interdependencies between the elements of the networks do not all have to be coded explicitly (cf. Pearl, 1988). Given a model that encodes only direct causal links, it is possible to use its structure to derive information about conditional independence and about indirect interdependencies (e.g., spurious correlations based on a common cause, or correlations between factors separated by multiple links in a causal chain). Previous research on the effect of prior knowledge on categorization has focused on demonstrations of the use of knowledge about direct, explicit causal relations. For example, Medin et al. (1987) showed that people are often insensitive to statistical correlations between features unless they can bring to bear explicit knowledge about a direct underlying causal relation. In a task involving construction of categories, people were more likely to recognize and make use of a correlation between dizziness and earache, for which they had prior causal knowledge, than a correlation between sore throat and skin rash, for which no such knowledge was readily available (for similar demonstrations see Murphy & Wisniewski, 1989; Pazzani, 1991; Wattenmaker, 1992; Wattenmaker et al., 1986). Our study extends such findings by demonstrating that people are also sensitive to within-category correlations that are indirect side effects of a more complex causal structure. For example, in Experiment 4 participants received instructions only about potential causal relations between the magnets and the orientation of individual compounds. Nonetheless, the learning data showed that this information sensitized them to the within-category correlation between effects implied by the structure of a common-cause model with a varying cause.

What is general about causal models, from the present perspective, is the generative use of information about a relatively small number of general structural properties to construct causal models tailored to particular learning situations. The properties on which we have focused in the present article are causal directionality, continuity of causal factors, and the variability of the causal variables. Candi-

dates that might be added to this list would include whether causes are expected to be deterministic or probabilistic, and the form of the function integrating the influence of multiple causes (e.g., summation vs. averaging). Although we believe the number of such general structural properties, which provide the building blocks for causal models, will prove to be relatively small, it should be readily apparent that the number of distinct causal models that could be constructed by forming combinations of such properties is enormous. It is this combinatorial richness that helps enable human causal induction to cope with the acquisition of novel causal structures, even without the benefit of prior knowledge about specific causal connections.

Causal-model theory thus leads us to reject the idea that human causal induction is based on a small number of relatively monolithic causal schemas. Nor would we expect our inductive apparatus to be tailored to purely syntactic distinctions such as linear separability (cf. Holland, Holyoak, Nisbett, & Thagard, 1986). In seeking theoretical generality within the theory-based view, some researchers have sought to identify ties between variations in linear separability and general content domains. For example, Wattenmaker (in press) reported a series of experiments on category sorting and learning that suggests that social categories (e.g., introverted vs. extroverted, active vs. passive, cautious vs. noncautious) generally lead to biases favoring LS categories, whereas object categories (e.g., furniture, animals, or vehicles) tend to be more compatible with NLS categories. Wattenmaker found biases of this sort when he used familiar materials; however, such differences might be attributed to use of specific world knowledge, rather than to more general differences between the structure of social and object categories. Wattenmaker also examined the acquisition of categories based on arbitrary social features. If social categories comprise a natural domain in which acquisition of LS structures is favored, then such structures should be learned relatively quickly given that a social context is established, even in the absence of more specific prior knowledge. However, Wattenmaker found no advantage for the LS arrangement in either of two experiments that used arbitrary social categories. Thus while Wattenmaker's (in press) results provided additional evidence that specific and explicit prior knowledge can mediate the ease of learning different category structures (cf. Wattenmaker et al., 1986), they provided no evidence that social or object domains per se serve as a basis for organizing causal learning.

It is possible that certain content domains are in fact strongly correlated with particular causal models (although in the absence of a systematic survey of the ecological frequency of different causal structures, there is no strong basis for positing such general correlations for social or object categories). If this was the case, then domain-related differences would not be caused by the different semantic contents of the learning material but rather by different underlying causal structures that are specified in a domain-general fashion. For example, the greater ease of NLS structures within the object domain may be due to the fact that common-cause models happen to occur more often in this domain than in other domains. Causal-model theory predicts that it is the underlying causal structure, rather than more superficial characteristics of the specific content domain, that primarily determines the representation of relations among features in causal categories.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409–429.

Bellingham, W. P., Gillette-Bellingham, K., & Kehoe, E. J. (1985). Summation and configuration in patterning schedules with the rat and rabbit. *Animal Learning & Behavior, 13,* 152–164.

Bolles, K. A (1989). *Structural equations with latent variables.* New York: Wiley.

Brehmer, B. (1969). Cognitive dependence on additive and configural cue–criterion relations. *American Journal of Psychology, 82,* 490–503.

Cartwright, N. (1989). *Nature's capacities and their measurement.* Oxford, England: Clarendon Press.

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review, 99,* 365–382.

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

Dawes, R. M. (1988). *Rational choice in an uncertain world.* San Diego, CA: Harcourt Brace Jovanovich.

Downing, C. J., Sternberg, R. J., & Ross, B. H. (1985). Multicausal inference: Evaluation of evidence in causally complex situations. *Journal of Experimental Psychology: General, 114,* 239–263.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). New York: Cambridge University Press.

Edgell, S. E., & Castellan, N. J., Jr. (1973). Configural effect in multiple-cue probability learning. *Journal of Experimental Psychology, 100,* 310–314.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin, 99,* 3–19.

Estes, W. K. (1986). Memory storage and retrieval processes in category learning. *Journal of Experimental Psychology: General, 115,* 155–174.

Gluck, M. A., & Bower, G. H. (1988a). Evaluating an adaptive network model of human learning. *Journal of Memory and Language, 27,* 166–195.

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

Gluck, M. A., Bower, G. H., & Hee, M. R. (1989). A configural-cue network model of animal and human associative learning. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 323–332). Hillsdale, NJ: Erlbaum.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction. Processes of inference, learning, and discovery.* Cambridge, MA: MIT Press.

Irzig, G., & Meyer, E. (1987). Causal modeling: New directions for statistical explanation. *Philosophy of Science, 54,* 495–514.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 276–296). New York: Appleton-Century-Crofts.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska Symposium on Motivation* (Vol. 15, pp. 192–238). Lincoln: University of Nebraska Press.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connec-

tionist model of category learning. *Psychological Review, 99,* 22–44.

Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior, 23,* 250–269.

Medin, D. L. (1975). A theory of context in discrimination learning. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 9, pp. 263–314). New York: Academic Press.

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 37–50.

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

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

Medin, D. L., & Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology, 20,* 158–190.

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

Mellers, B. A. (1980). Configurality in multiple-cue probability learning. *American Journal of Psychology, 93,* 429–443.

Melz, E. R., Cheng, P. W., Holyoak, K. J., & Waldmann, M. R. (1993). Cue competition in human categorization: Contingency or the Rescorla–Wagner learning rule? Comments on Shanks (1991). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1398–1410.

Minsky, M., & Papert, S. (1969). *Perceptrons: An introduction to computational geometry.* Cambridge, MA: MIT Press.

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

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

Nakamura, G. V. (1985). Knowledge-based classification of ill-defined categories. *Memory & Cognition, 13,* 377–384.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 104–114.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. T. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition, 22,* 352–369.

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

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Mateo, CA: Morgan Kaufmann.

Peng, Y., & Reggia, J. A. (1990). *Abductive inference-models for diagnostic problem-solving.* New York: Springer-Verlag.

Reichenbach, H. (1956). *The direction of time.* Berkeley and Los Angeles: University of California Press.

Rescorla, R. A. (1972). "Configural" conditioning in discrete-trial bar pressing. *Journal of Comparative and Physiological Psychology, 79,* 307–317.

Rescorla, R. A. (1973). Evidence for the "unique stimulus" account of configural conditioning. *Journal of Comparative and Physiological Psychology, 85,* 331–338.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world.* Princeton, NJ: Princeton University Press.

Shanks, D. R. (1991). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 433–443.

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 21, pp. 229–261). New York: Academic Press.

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75*(13), Whole No. 517.

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (pp. 49–72). Hillsdale, NJ: Erlbaum.

van Hamme, L. J., Kao, S. F., & Wasserman, E. A. (1993). Judging intervent relations: From cause to effect and from effect to cause. *Memory & Cognition, 21,* 802–808.

Waldmann, M. R., & Holyoak, K. J. (1990). Can causal induction be reduced to associative learning? *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 190–197). Hillsdale, NJ: Erlbaum.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General, 121,* 222–236.

Wasserman, E. A. (1990). Attribution of causality to common and distinctive elements of compound stimuli. *Psychological Science, 1,* 298–302.

Wattenmaker, W. D. (1991). Learning modes, feature correlations, and memory-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 908–923.

Wattenmaker, W. D. (1992). Relational properties and memory-based category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1125–1138.

Wattenmaker, W. D. (1993). Incidental concept learning, feature frequency, and correlated properties. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 203–222.

Wattenmaker, W. D. (in press). Knowledge structures and linear separability: Integrating information in object and social categorization. *Cognitive Psychology.*

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

Widrow, G., & Hoff, M. E. (1960). Adaptive switching circuits. *Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record, 4,* 96–194.

Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science, 18,* 221–281.

Received April 6, 1994
Revision received December 16, 1994
Accepted December 20, 1994 ■