

A Causal-Model Theory of Conceptual Representation and Categorization

Bob Rehder
New York University

This article presents a theory of categorization that accounts for the effects of causal knowledge that relates the features of categories. According to causal-model theory, people explicitly represent the probabilistic causal mechanisms that link category features and classify objects by evaluating whether they were likely to have been generated by those mechanisms. In 3 experiments, participants were taught causal knowledge that related the features of a novel category. Causal-model theory provided a good quantitative account of the effect of this knowledge on the importance of both individual features and interfeature correlations to classification. By enabling precise model fits and interpretable parameter estimates, causal-model theory helps place the theory-based approach to conceptual representation on equal footing with the well-known similarity-based approaches.

For the last several decades, research on the topic of categorization has focused on the problem of learning new categories via examples of category members, that is, from empirical observations. The result has been a host of categorization models that are based on representational ideas such as central prototypes, stored exemplars, and variabilized rules, and on processing principles such as similarity, that have considerable explanatory power and experimental support. More recently, the influence of the prior “theoretical” knowledge that learners often contribute to their representations of categories has also been a topic of study (Carey, 1985; Keil, 1989; Murphy & Medin, 1985; Schank, Collins, & Hunter, 1986). For example, people not only know that birds have wings and that they can fly and build nests in trees, but also that birds build nests in trees *because* they can fly, and fly *because* they have wings. Many people even believe that morphological features of birds such as wings are ultimately caused by the kind of DNA that birds possess. However, in comparison with the development of models accounting for the effects of empirical observations, there has been relatively little development of formal models to account for the effects of such prior knowledge (although see Heit, 1994; Heit & Bott, 2000; Pazzani, 1991; Rehder & Murphy, in press; Sloman, Love, & Ahn, 1998).

The purpose of this article is to present a theory of categorization that accounts for the effects of theoretical knowledge, particularly causal knowledge, that interrelates or links the features of many categories that people possess. According to causal-model theory, people’s knowledge of many categories includes not just a representation of a category’s features but also an explicit representation of the causal mechanisms that people believe link those

features (Rehder, 1999; Waldmann, Holyoak, & Fratianne, 1995). Further, according to this theory, people use causal models to determine a new object’s category membership.

In this article, causal-model theory is applied to two outstanding problems in the domain of categorization research. The first problem concerns determining the importance, or *weight*, that individual features have on establishing category membership. Since the popularization of the notion of probabilistic categories in the 1970s, it has usually been assumed that features of a category vary regarding their influence on category membership (Hampton, 1979; Rosch, 1973; Rosch & Mervis, 1975; Smith & Medin, 1981). Indeed, formal models of categorization have formalized the manner in which a feature’s weight is influenced by its perceptual saliency (Lamberts, 1995, 1998) and by the frequency with which it appears in category members and nonmembers (Nosofsky, 1986; Reed, 1972; Rosch & Mervis, 1975; Shepard, Hovland, & Jenkins, 1961). However, these models do not account for the fact that feature weights are also determined by a categorizer’s domain theories. For instance, Medin and Shoben (1988) have found that *straight bananas* are rated as better members of the category *bananas* than *straight boomerangs* are of the category *boomerangs*, a result they attribute to the default feature *curved* occupying a more theoretically central position in the conceptual representation of *boomerang* as compared with *banana* (also see Kaplan & Murphy, 2000; Murphy & Allopenna, 1994; Pazzani, 1991; Wattenmaker, Dewey, Murphy, & Medin, 1986). Keil (1989) found that second and fourth graders judged that animals with transformed perceptual features retained their category membership (e.g., a raccoon that is made skunk-like by being dyed black, painted with a white stripe, and given a “sac of super smelly yucky stuff” is still a raccoon), a result Keil attributed to the children’s belief that the theoretically central (albeit hidden) features of the kind important for category membership had been left unaltered (also see Gelman & Wellman, 1991; Rips, 1989). Rehder and Hastie (2001) showed that category features involved in many causal relationships have more influence on category judgments than other features (for related results, see Ahn, 1998; Ahn, Kim, Lassaline, & Dennis, 2000; Sloman et al., 1998).

I thank Patricia Berretty, Seth Chin-Parker, Reid Hastie, Evan Heit, Greg Murphy, Brian Ross, Cindy Sifonis, and two anonymous reviewers for their comments on earlier versions of this article. Support for this research was provided by Grants SBR-9816458 and SBR 97-20304 from the National Science Foundation and Grant R01 MH58362 from the National Institute of Mental Health.

Correspondence concerning this article should be addressed to Bob Rehder, Department of Psychology, New York University, 6 Washington Place, New York, New York 10003. E-mail: bob.rehder@nyu.edu

The second outstanding problem concerns the effect of specific configurations of features on category membership decisions. That is, above and beyond the influence of individual features, particular combinations of features displayed by an object may be seen as good evidence for or against category membership. For example, previous research has found that sensitivity to interfeature correlations can emerge when people categorize by analogy to previously observed category members (Medin & Schaffer, 1978; Nosofsky, 1986; also see Anderson & Fincham, 1996; Thomas, 1998). However, current categorization theories do not account for how domain theories also influence which feature configurations make for acceptable category members. For example, Wattenmaker et al. (1986) elicited sensitivity to particular correlations of features (between working indoors vs. outdoors, and working all year vs. working in the nonwinter months) by reminding participants of the existence of both indoor and outdoor painters. Murphy and Wisniewski (1989) found that whether categorization judgments were affected by feature combinations depended on whether those combinations were expected on the basis of prior knowledge rather than on whether they had been previously observed in category exemplars (also see Chapman & Chapman, 1967, 1969). Wisniewski (1995) found that certain objects were better examples of the category *captures animals* when they possessed novel combinations of features that were useful (e.g., *contains peanuts* and *caught a squirrel*) versus when they did not (e.g., *contains acorns* and *caught an elephant*; also see Malt & Smith, 1984; Rehder & Ross, 2001).

The central claim of causal-model theory is that sensitivity to features and specific configurations of features can often be attributed to the presence of causal knowledge about a category. According to causal-model theory, features and configurations of features influence judgments of category membership to the extent they are likely to be produced, or *generated*, by a category's causal laws. For example, assuming that people believe that bird DNA causes wings, which cause flying, which causes nests in trees, they should also believe (all things being equal) that features that are more directly caused by bird DNA (e.g., having wings) are more reliably generated than those that are more indirectly caused (e.g., building nests in trees). As a result, directly caused features should be viewed as occurring more frequently among category members (and hence should be weighed more heavily in judgments of category membership). In addition, causal-model theory also claims that combinations of features are important pieces of evidence for category membership to the extent that they are jointly consistent or inconsistent with the category's causal knowledge. For example, an animal that does not fly and yet still builds nests in trees might be considered a less plausible bird (how did the nest get into the tree?) than an animal that does not fly and builds nests on the ground (e.g., an ostrich), even though the former animal has more features that are typical of birds.

This article presents causal-model theory's formal account of how causal knowledge influences the importance of features and specific configurations of features in judgments of category membership. Previous investigators have suggested that the theoretical knowledge that people possess about categories may be conceived of as "constraining or even generating properties" that we observe in category members (Medin & Ortony, 1989, p. 185). Causal-model theory formalizes this proposal by specifying the feature weights that can be expected given a category's causal laws and

also extends it by additionally specifying the pattern of expected interfeature correlations. As will be demonstrated, causal-model theory yields a precise, quantitative account of both the differences in feature weights and the importance of feature configurations induced by knowledge of a category's causal laws that emerges when categorization decisions are being made.

Because people's theoretical knowledge of many natural categories is likely to be complex and to vary considerably from person to person, this study used novel categories to provide experimental control over the knowledge associated with a category. Table 1 presents an example of the features and causal relationships for one of the novel categories, *Lake Victoria Shrimp*. Lake Victoria Shrimp were described to participants as possessing four binary features and three causal relationships among those features. The causal links were arranged in a chain pattern such that the first feature causes the second feature, which causes the third feature, which causes the fourth. The knowledge associated with categories such as *Lake Victoria Shrimp* was intended to be a simplified analogue of real-world category knowledge, such as the knowledge that bird DNA causes wings, which cause flying, which causes nests in trees.

In the following section, I present causal-model theory, which includes proposals regarding the representation of causal knowledge and a method for computing evidence that an object is a category member in light of that knowledge. The predictions of causal-model theory regarding feature weights and interfeature correlations will be shown to be based on the structure of causal knowledge rather than its content (i.e., the details of the causal mechanisms), an assumption I tested by using five novel categories in addition to *Lake Victoria Shrimp*.

Causal-Model Theory

The central claim of causal-model theory is that people's knowledge of categories includes not just category features but also the representation of causal mechanisms that link those features. Figure 1A demonstrates a simple causal model in which one feature, *C*, is depicted as the cause of a second feature, *E*. Causal models are specialized instances of Bayesian networks. Bayesian networks consist of nodes that represent variables and directed edges that can be interpreted as representing direct causal relationships between variables. In this study, the variables are binary category attributes that represent whether a feature is present or absent (1 = present; 0 = absent). (See Glymour, 1998; Jordan, 1999; and Pearl, 1988, 2000, for complete descriptions of Bayesian networks.)

A Bayesian network represents the fact that an effect variable is causally influenced by its immediate parents (technically, that the effect variable's probability distribution is conditionally independent of any nondescendent when the state of the parent variables is known). However, by itself, the network conveys no information regarding the details of the causal relationships that link directly connected variables in a network. In contrast, causal-model theory makes specific assumptions regarding how people conceive of causal relationships between binary variables. Specifically, it assumes that people view features as being linked by probabilistic causal mechanisms. For example, in this article it is assumed that when a cause feature is present (e.g., $C = 1$ in Figure 1A) it enables the operation of a causal mechanism that will, with some

Table 1
Features and Causal Relationships for the Lake Victoria Shrimp Experimental Category

Features and causal relationships	Descriptions
F ₁	High amounts of ACh neurotransmitter.
F ₂	Long-lasting flight response.
F ₃	Accelerated sleep cycle.
F ₄	High body weight.
F ₁ → F ₂	A high quantity of the ACh neurotransmitter causes a long-lasting flight response. The duration of the electrical signal to the muscles is longer because of the excess amount of neurotransmitter.
F ₂ → F ₃	A long-lasting flight response causes an accelerated sleep cycle. The long-lasting flight response causes the muscles to be fatigued, and this fatigue triggers the shrimp's sleep center.
F ₃ → F ₄	An accelerated sleep cycle causes a high body weight. Shrimp habitually feed after waking, and shrimp on an accelerated sleep cycle wake three times a day instead of once.

probability, bring about the presence of the effect feature (e.g., $E = 1$ in Figure 1A). When the cause feature, C , is absent it is assumed that it has no causal influence on the effect, E .

The second major claim of causal-model theory is that categorizers make classification decisions by estimating how likely an exemplar is to have been generated by a category's causal model. The causal model of Figure 1A has three parameters, each of which are probabilities in the range from 0 to 1. Together with elementary probability theory, these parameters specify the likelihood that the model will generate any combination of the two features C and E . Parameter c represents the probability that feature C will be present. Parameter m represents the probability that the probabilistic mechanism that links C and E will successfully operate (that is, will bring about the presence of E) when C is present. Parameter b represents the probability that E will be present even when it is not brought about by C . Parameter b can be interpreted as the probability that E is brought about by unspecified background causes other than C .

Table 2 presents the likelihoods that the causal model of Figure

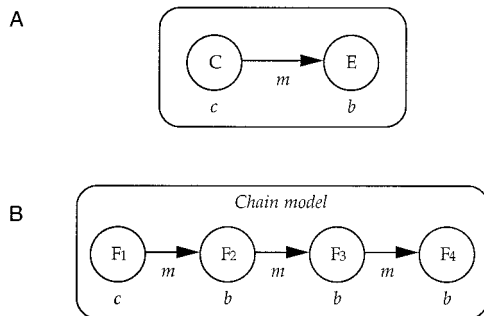


Figure 1. A: A simple causal model with two binary features and one causal relationship. B: A chain causal model.

Table 2
Likelihood Equations for a Causal Model With Two Binary Features and One Causal Relationship

Exemplar (E)	$L(E; c, m, b)$	$L(E; .50, .80, .20)$
00	$(1 - c)(1 - b)$.40
01	$(1 - c)(b)$.10
10	$(c)[(1 - m)(1 - b)]$.08
11	$(c)(m + b - mb)$.42

Note. L = the likelihood of exemplar E as a function of c , m , and b ; c = the probability of the cause feature; m = the strength of the causal mechanism; b = the strength of the background mechanism.

1A will generate the four possible combinations of C and E in terms of the parameters c , m , and b . The probability that C and E will both be absent, that is, $P(\sim C \sim E)$, also referred to as $P(00)$, is the probability that C is absent $(1 - c)$ times the probability that E is not brought about by any background causes $(1 - b)$. Note that parameter m is not involved in this likelihood because it is assumed that the causal mechanism relating C and E only potentially operates when C is present. The probability that C is absent but E is present, that is, $P(\sim CE)$ or $P(01)$ is $1 - c$ times the probability that E is brought about by some background cause, b . The probability that C is present but E is absent, that is, $P(C \sim E)$ or $P(10)$ is c times the probability that E is not brought about by either the causal mechanism or the background cause, $(1 - m)(1 - b)$. Finally, the probability that C and E are both present, that is, $P(CE)$ or $P(11)$ is c times the probability that E is brought about by the causal mechanism or brought about by the background cause $(m + b - mb)$. Note that Table 2 represents a proper probability function because $P(00) + P(01) + P(10) + P(11) = 1$ for any values of c , m , and b in the range from 0 to 1.

Parameters b and m correspond to probabilistic versions of the familiar notions of the necessity and sufficiency of causal relations. *Causal necessity* (an effect is always accompanied by its cause) obtains when $b = 0$; *causal sufficiency* (a cause is always accompanied by its effect) obtains when $m = 1$. Thus, in causal-model theory deterministic causality appears as a limiting case. Table 2 presents the likelihoods for a case of near necessity and near sufficiency, namely, $m = .80$ and $b = .20$. In this case, exemplars that represent violations of necessity and sufficiency (01 and 10) are unlikely to be generated (probabilities of .10 and .08, respectively).

Causal models may include an unlimited number of variables causally connected in any pattern that avoids cycles. This research investigates the characteristics of the chain causal model (see Figure 1B) in which the binary category feature, F_1 , causes F_2 , which in turn causes F_3 , which in turn causes F_4 . In the chain model, it is assumed that the three causal mechanisms interconnecting the four features operate independently, and each with probability m . It is also assumed that the background causes of F_2 , F_3 , and F_4 operate independently and each with probability b . (More discussion about these assumptions follows.) The likelihood equations for the 16 possible combinations of values for F_1 , F_2 , F_3 , and F_4 as a function of parameters c , m , and b can be generated by recursively applying the equations from Table 2. These equations are presented in Table 3. For example, the probability of exemplar 0110 (that is, F_2 and F_3 present, F_1 and F_4 absent) being generated

Table 3
Likelihood Equations for the Chain Model

Exemplar (E)	$L_{Chain}(E; c, m, b)$
0000	$(1 - c)(1 - b)(1 - b)(1 - b)$
0001	$(1 - c)(1 - b)(1 - b)(b)$
0010	$(1 - c)(1 - b)(b)[(1 - m)(1 - b)]$
0100	$(1 - c)(b)[(1 - m)(1 - b)](1 - b)$
1000	$(c)[(1 - m)(1 - b)](1 - b)(1 - b)$
0011	$(1 - c)(1 - b)(b)(m + b - mb)$
0101	$(1 - c)(b)[(1 - m)(1 - b)](b)$
1001	$(c)[(1 - m)(1 - b)](1 - b)(b)$
0110	$(1 - c)(b)(m + b - mb)[(1 - m)(1 - b)]$
1010	$(c)[(1 - m)(1 - b)](b)[(1 - m)(1 - b)]$
1100	$(c)(m + b - mb)[(1 - m)(1 - b)](1 - b)$
0111	$(1 - c)(b)(m + b - mb)(m + b - mb)$
1011	$(c)[(1 - m)(1 - b)](b)(m + b - mb)$
1101	$(c)(m + b - mb)[(1 - m)(1 - b)](b)$
1110	$(c)(m + b - mb)(m + b - mb)[(1 - m)(1 - b)]$
1111	$(c)(m + b - mb)(m + b - mb)(m + b - mb)$

Note. L = the likelihood of exemplar E as a function of $c, m,$ and b ; c = the probability of the cause feature; m = the strength of the causal mechanism; b = the strength of the background mechanism.

by a chain causal model is the probability that F_1 is absent $(1 - c)$, times the probability that F_2 is brought about by its background cause (b) , times the probability that F_3 is brought about by F_2 or its background cause $(m + b - mb)$, times the probability that F_4 is brought about by neither F_3 nor its background cause, that is, $(1 - m)(1 - b)$.

How the likelihoods generated by causal models are translated into categorization judgments depends on the nature of the decision task presented to participants. For example, when there are two candidate categories to which an object might belong, the likelihoods from each category’s causal model may be combined according to Luce’s (1963) choice axiom to predict choice probabilities. For example, the probability that an exemplar, E , will be classified into Category A versus Category B would be given by $P(A|E) = L_A(E)/[L_A(E) + L_B(E)]$, where L_A and L_B are the likelihoods that E would be generated by A’s and B’s causal models, respectively. However, in this study undergraduates were first taught a single novel category and then were asked to rate the category membership of a number of exemplars. The assumption is that the rating assigned to an exemplar, E , will be equal to its likelihood with respect to the single category’s causal model, scaled by a constant, K ,

$$\text{Rating}(E) = KL_{Chain}(E; c, m, b). \tag{1}$$

Causal Models and Derived Empirical Statistics

The central question taken up in the current article concerns the effect of causal knowledge on the importance of features and interfeature correlations on categorization decisions. According to causal-model theory, the sensitivity to features and interfeature correlations emerges from categorizers assessing how likely an exemplar is to have been generated by a category’s causal laws. This section demonstrates how to derive the feature probabilities and interfeature correlations that are generated by a causal model and are therefore predictive of category membership.

Feature Probabilities

By definition, the probability of cause features, for example, $P(C)$ in Figure 1A or $P(F_1)$ in Figure 1B, is simply its associated c parameter. In contrast, the probability of an effect feature is a function of whether it occurs on its own (as represented by its b parameter) or is brought about by one of its causes. For an effect feature F_i ($i = 2, 3,$ or 4) in the chain causal model of Figure 1B, Equation 2 expresses the probability that F_i is present (P_i) as the probability that it is brought about by the causal mechanism (the probability that the cause feature is present, P_{i-1} , times the probability that the causal mechanism operates, m) or by some background cause (b) ,

$$P_i = P_{i-1}m + b - P_{i-1}mb, \tag{2}$$

which simplifies to

$$P_i = P_{i-1}m(1 - b) + b. \tag{3}$$

Equation 3 indicates that the probability of F_i increases both as the probability of its cause F_{i-1} (P_{i-1}) increases and as the strength of the causal mechanism linking F_{i-1} and F_i (m) increases. To illustrate, Figure 2 presents the probabilities of the four features for a chain causal model for various values of $c, m,$ and b . In Figure 2A–E, Parameter c is varied from 1.00 to .75 to .50; Parameter b is varied from .25 to .50; and Parameter m is varied over the values 0.00, .10, .40, and .80.

Figure 2 demonstrates two aspects of the effect of causal mechanisms on feature probabilities as expressed in Equation 2. First, the presence of a causal mechanism always increases the probability of an effect. When there are no causal mechanisms between features (i.e., $m = 0$), the probability of $F_2, F_3,$ and F_4 is equal to parameter b . However, the introduction of causal mechanisms between features makes the effect features $F_2, F_3,$ and F_4 more probable than they would be otherwise: In each graph, $P_2, P_3,$ and P_4 increase as m increases.

Second, Figure 2 indicates that the probability of effects decreases as the probability of their causes decreases. For example, when the root cause is likely, features will become less probable as one moves down the chain. This situation is illustrated by case $c = 1.00, b = .25, m = .40$ of Figure 2A, where $P_1, P_2, P_3,$ and P_4 are 1.00, .55, .41, and .37, respectively. This case corresponds to the bird example presented earlier: If people believe that bird DNA causes wings, which cause flying, which causes nests in trees, they should also believe that features that are more directly caused by DNA, such as having wings, will be more probable than those features that are more indirectly caused by DNA, such as building nests in trees.

Although analytic expressions for feature probabilities such as Equation 3 can be derived for any parameterized causal model, feature probabilities can be calculated from any function that specifies exemplar probabilities by summing the probabilities of those exemplars that possess the features. For example, for a category that possesses four binary features, the probability that F_4 is present is the sum of the probabilities of the exemplars that possess F_4 : $P_4 = P(0001) + P(0011) + P(0101) + P(0111) + P(1001) + P(1011) + P(1101) + P(1111)$.

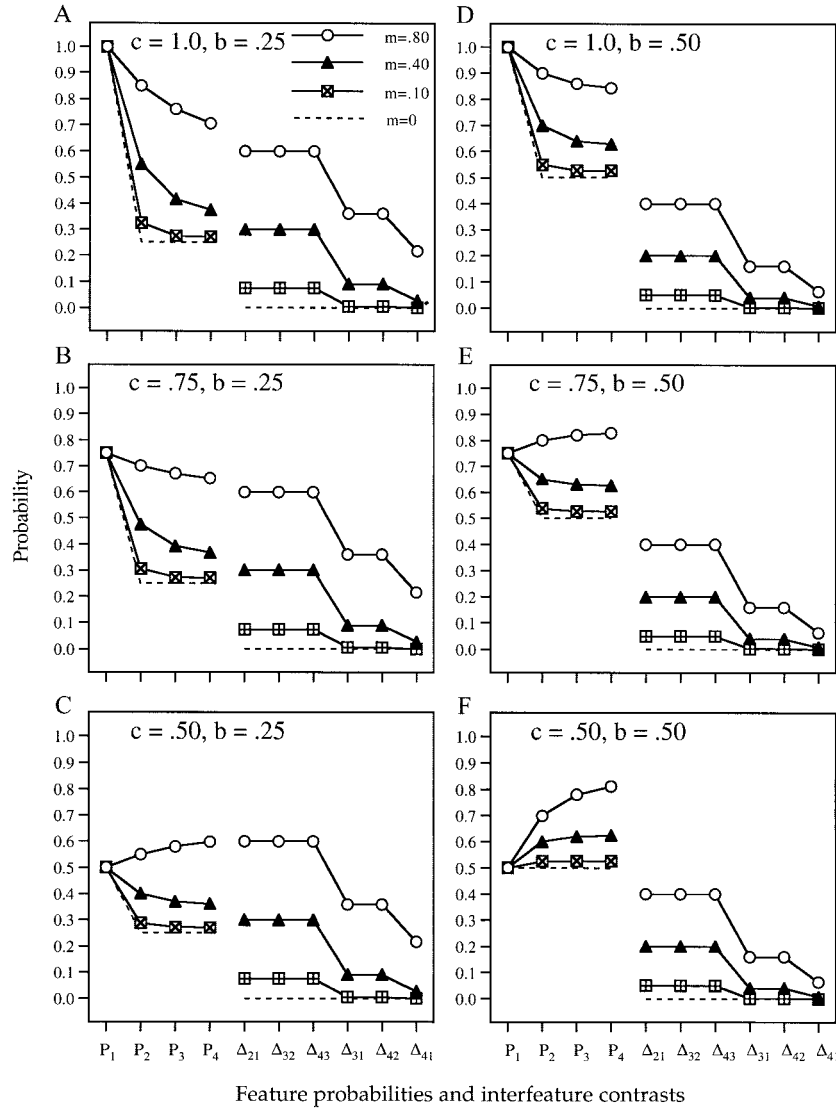


Figure 2. Feature weights and interfeature correlations for the chain causal model for various values of c , m , and b .

Interfeature Correlations

According to causal-model theory, causal relationships among features also imply a certain pattern of interfeature correlations. A convenient measure of feature co-occurrence that has been used in research on the induction of causal relationships is the probabilistic contrast (e.g., Cheng & Novick, 1990). In this view, the strength of covariation between causes and effects is the probability of the effect in the presence of the cause minus the probability of the effect in the absence of the cause. That is, the probabilistic contrast between a cause feature, X , and an effect feature, Y , is,

$$\Delta_{YX} = P(Y|X) - P(Y|\sim X) \tag{4}$$

For a chain causal model, the contrast between features directly connected by causal relationships increases as the strength of the causal mechanism (as represented by parameter m) increases, and

as the probability that the background mechanism brings about the effect (as represented by parameter b) decreases, as shown in Equation 5:

$$\Delta_{i,i-1} = m(1 - b), \tag{5}$$

for $i = 2, 3$, or 4 . In general, the probabilistic contrast Δ between any pair of features can be calculated from any function that specifies exemplar probabilities. For example, for the chain causal model, $\Delta_{41} = P(F_4|F_1) - P(F_4|\sim F_1)$, and $P(F_4|F_1)$, and $P(F_4|\sim F_1)$ can each be calculated directly from the likelihood equations of Table 3: $P(F_4|F_1) = [P(1001) + P(1011) + P(1101) + P(1111)]/P_1$, and $P(F_4|\sim F_1) = [P(0001) + P(0011) + P(0101) + P(0111)]/(1 - P_1)$.

Figure 2 presents the contrasts for all six feature pairs for the chain causal model for various values of m , and b . Two things should be noted. First, as expected given Equation 5, the value of

Δ increases as the m parameter increases (reflecting the greater strength of the interfeature causal mechanisms), and as the b parameter decreases (reflecting the weaker strength of alternative background causes that might bring about the effects). Second, the interfeature contrasts are strongest for those feature pairs that are directly connected by causal mechanisms (F_1 and F_2 , F_2 and F_3 , and F_3 and F_4), weaker for those pairs that are separated by one intervening variable (F_1 and F_3 , and F_2 and F_4), and weakest for those pairs that are separated by two intervening variables (F_1 and F_4). This fact is in accordance with our intuitions that variables will be less strongly correlated as their causal connectedness becomes more distant.

In summary, the examples presented in Figure 2 illustrate how a chain causal model predicts a distinctive pattern of feature weights and interfeature correlations. Causal-model theory demonstrates how the introduction of causal mechanisms between features simultaneously affects both the weights associated with features (Equation 3) and the pattern of interfeature correlations (Equation 5). In other words, feature weights and interfeature correlations are an emergent property of a procedure that categorizes according to whether an object was likely to have been generated by a category's causal laws.

In the following three experiments, participants first learned about a novel category such as *Lake Victoria Shrimp* and then generated categorization ratings for specific exemplars. To test the claim that an exemplar's rating should be proportional to the likelihood that it was generated by the category's causal laws, each participant's ratings were fit to a chain causal model by estimating the values of parameters c , m , and b that maximized the fit between the ratings and the model likelihoods. (Parameter K , which rescales the ratings according to Equation 1, was also estimated for each participant.) The three experiments differed regarding the feature base-rate information that participants received. In Experiment 1, participants were told that all features occurred 75% of the time. In Experiment 2, participants were provided with no base-rate information. To investigate performance with categories that possess an underlying, causally potent defining feature or "essence" (e.g., bird DNA), I designed Experiment 3 so that participants were told that the root cause (F_1) occurred in 100% of category members, whereas no base-rate information about F_2 , F_3 , and F_4 was provided.

Experiment 1

Method

Materials. Six novel categories were constructed: two biological kinds (*Keohoe Ants*, *Lake Victoria Shrimp*), two nonliving natural kinds (*Myastars*, *Meteoric Sodium Carbonate*), and two artifacts (*Romanian Rogos*, *Neptune Personal Computers*). The features and the causal relationships for each category are presented in the Appendix.

The base rate for each feature was described as 75%. For example, 75% of *Lake Victoria Shrimp* have high body weight, whereas 25% have normal body weight. Participants in the chain condition also learned of three interfeature causal relationships arranged as in Figure 1B. This description consisted of one sentence indicating the cause and effect features (e.g., *A high quantity of ACh neurotransmitter causes a long-lasting flight response*) and then one or two sentences briefly describing the causal mechanism (e.g., *The duration of the electrical signal to the muscles is longer because of the excess amount of neurotransmitter*).

Procedure. Experimental sessions were conducted by computer. Participants first studied several screens of information about their assigned category at their own pace. All participants were presented with the cover story and the category's features and their base rates; participants in the chain condition were additionally presented with a description of three causal relationships and a diagram depicting those relationships like that in Figure 1B. When ready, participants took a multiple-choice test of the information they had just studied. Participants could request help, in which case the computer re-presented the information about the category. Participants were required to retake the test until they committed 0 errors and made 0 requests for help.

Participants then performed three tasks counterbalanced for order: a categorization task, a property-induction task, and a similarity-rating task. The results from the property-induction and similarity-rating task are unrelated to the theoretical issues raised in this article and are omitted. During the categorization task, participants rated the category membership of 48 exemplars, consisting of all 16 possible objects that can be formed from 4 binary attributes and the 8 single-attribute exemplars, each presented twice. For example, those participants assigned to learn the *Lake Victoria Shrimp* category were shown a shrimp that possessed *high amounts of the ACh neurotransmitter*, *a normal flight response*, *accelerated sleep cycle*, and *normal body weight*, and were asked, *is this a Lake Victoria Shrimp?* The attribute values of each to-be-rated exemplar were listed in order (dimensions 1–4) on the computer screen. The list of attribute values for single-attribute exemplars contained ??? for the three unknown attributes. The order of the 48 exemplars was randomized for each participant.

Participants entered their rating by using the left and right arrow keys to move a bar along a response scale to a position that reflected their confidence that the exemplar was a category member. The two ends of the scale were labeled *definitely not an X* and *definitely an X*, where X was the name of the category. The response bar could be set to 21 distinct positions. Responses were scaled into a number in the range 0–100.

Participants. Seventy-two University of Colorado at Boulder undergraduates received course credit for their participation in this experiment. They were assigned in equal numbers to the chain and control conditions, and to one of the six experimental categories.

Results

Category membership ratings for the 16 test exemplars averaged over participants in the chain and control conditions are presented in Table 4. As expected, the 75% feature base-rate information with which participants were provided had an overall effect on category membership ratings, as exemplars that possessed more features received a higher categorization rating. For example, the average rating for the exemplar missing all four features (0000) was 38.9 as compared to 90.0 for the exemplar possessing all four features (1111).

In addition, the presence of interfeature causal relationships in the chain condition affected the categorization ratings relative to the control condition. For instance, exemplars 0101 and 1010 were given significantly lower ratings in the chain condition (39.6 and 45.1, respectively) than in the control condition (59.7 and 57.3). Presumably, the lower ratings for 0101 and 1010 arose because those exemplars violate all three causal laws. Exemplar 0101 violates all three laws because Features F_2 and F_4 are present even though their causes (F_1 and F_3 , respectively) are absent, and F_3 is absent even though its cause (F_2) is present. Analogously, exemplar 1010 violates all three causal relationships because F_2 and F_4 are absent even though their causes (F_1 and F_3 , respectively) are present, and F_3 is present even though its cause (F_2) is absent. The

Table 4
Exemplar Classification Ratings From Experiments 1 and 2

Exemplar	Experiment 1				Experiment 2			
	Chain		Control		Chain		Control	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
0000	38.4	5.8	39.4	4.9	68.0	5.4	70.7	3.3
0001	33.8	3.3	47.4	3.7	38.4	3.6	67.0	3.1
0010	33.8	3.6	48.1	3.9	32.9	3.3	65.6	3.5
0100	31.7	3.5	47.8	3.8	29.0	3.3	66.0	3.4
1000	44.9	3.9	44.0	3.8	31.9	3.9	67.0	3.6
0011	49.0	3.3	55.4	3.0	37.8	3.4	67.1	3.1
0101	39.6	3.0	59.7	2.7	26.9	3.1	66.5	3.5
0110	42.1	3.0	57.6	2.9	28.5	3.2	65.6	3.3
1001	48.8	3.3	57.9	3.2	30.7	4.0	68.0	2.9
1010	45.1	3.4	57.4	2.8	28.1	3.8	67.6	3.3
1100	53.7	3.0	58.3	2.6	37.1	4.1	69.9	2.8
0111	61.6	4.1	69.4	3.0	40.5	4.2	67.6	3.1
1011	62.8	3.8	69.4	2.4	36.5	4.3	67.2	2.9
1101	62.9	4.0	71.6	2.3	41.5	4.4	70.2	2.9
1110	71.4	2.8	71.7	2.5	49.7	5.1	72.2	2.7
1111	93.2	2.0	84.7	2.6	91.6	2.9	75.6	2.7

violations of causal relationships in exemplars 0101 and 1010 meant that their ratings were not significantly higher than the rating received by exemplar 0000 (41.6) even though they displayed two characteristic features whereas 0000 displayed none. In contrast, exemplar 1111 received a significantly higher rating in the chain versus control condition (93.2 vs. 84.7), presumably because in the chain condition exemplar 1111 confirms all three causal links (cause and effect pairs are all present). The categorization ratings received by exemplars 0000, 0101, 1010, and 1111 are presented in Figure 3A as a function of experimental condition.

These conclusions were supported by statistical analysis. An initial analysis of variance (ANOVA) that included the particular category (six levels) as a factor revealed that the category had no effect on categorization ratings and did not interact with the other experimental factors. Hence, a two-way ANOVA with condition (2 levels: chain vs. control) and exemplar (16 levels) with repeated measures on the second factor was conducted. Consistent with the claim that participants were using the 75% feature base-rate information, there was a significant effect of exemplar, $F(15, 1050) = 47.4$, $MSE = 293.6$, $p < .0001$. In addition, there was a significant main effect of condition, $F(1, 70) = 8.17$, $MSE = 2227.6$, $p < .01$, and a significant interaction between experimental condition and exemplar, $F(15, 1050) = 3.42$, $MSE = 293.6$, $p < .0001$, reflecting the fact that the pattern of ratings given to exemplars differed in the chain and control conditions.

Feature probabilities. A central question asked in this article concerns the effects of causal knowledge on the importance of individual features in categorization. To address this issue, each participant's set of category membership ratings for the 16 test exemplars was treated as a set of probabilities (i.e., likelihoods) by rescaling the ratings so that they summed to 1. The procedure for computing feature probabilities from a set of likelihoods described earlier was then carried out for each participant. For example, the probability of P_1 for a participant was computed by summing the likelihoods for each exemplar that possessed F_1 (i.e., 1000, 1001, 1010, 1011, 1100, 1101, 1110, and 1111).

The four feature probabilities, P_1 , P_2 , P_3 , and P_4 , averaged over participants in the chain and control groups are displayed in Figure 3B. In the chain condition the values of P_1 , P_2 , P_3 , and P_4 were .600, .565, .567, and .558 as compared to .554, .562, .553 and .554 in the control condition. The fact that these feature probabilities are all greater than .50 for both the chain and control groups reflects that in both conditions the presence of a feature increased categorization ratings, and the absence of a feature decreased it. The approximately equal feature probabilities in the control condition indicated that each feature was equally important to the ratings. In contrast, in the chain condition the greater probability of the first feature F_1 indicated that it had more influence on the ratings than the other three features. This result obtained despite the fact that chain participants were told that all features were equally probable (i.e., 75%).

A two-way ANOVA of these feature probabilities with condition (chain vs. control) and feature (1, 2, 3, and 4) as factors with repeated measures on the last factor revealed a significant interaction between condition and feature, $F(3, 210) = 4.33$, $MSE = 0.00169$, $p < .01$, indicating that the pattern of feature probabilities differed in the two conditions. A separate analysis of the chain condition confirmed that the probability of F_1 was significantly greater than the average probability of F_2 , F_3 , and F_4 , $F(1, 35) = 6.66$, $MSE = 0.00743$, $p < .05$. The probabilities of F_2 , F_3 , and F_4 were not significantly different from one another.

Interfeature contrasts. To evaluate the effects of the chain condition in terms of interfeature correlations, or contrasts, I treated each participant's set of ratings for the 16 test exemplars as a set of likelihoods as in the previous section. The procedure for computing feature contrasts from a set of likelihoods described earlier was then carried out. The six interfeature contrasts, Δ_{21} , Δ_{32} , Δ_{43} , Δ_{31} , Δ_{42} , and Δ_{41} , averaged over participants are displayed in Figure 3B for the chain and control groups. Figure 3B indicates that in the chain condition the contrasts between those features directly connected in the causal chain (that is, Δ_{21} , Δ_{32} , and Δ_{43}) were .047, .054, and .056, respectively. Each of these

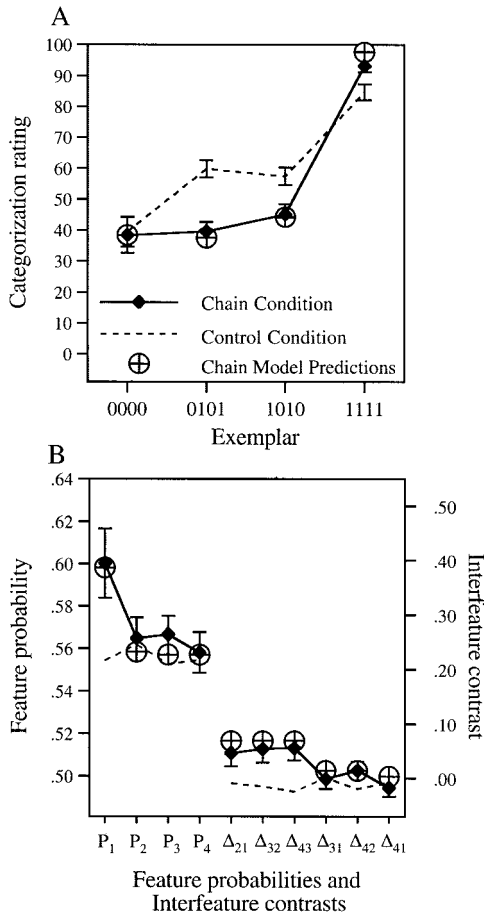


Figure 3. Results from the chain and control conditions of Experiment 1. A: Categorization ratings for selected exemplars. B: Derived feature probabilities and interfeature correlations. Predictions of the chain model are shown in each panel.

probabilities are greater than 0, reflecting that the joint presence (or the joint absence) of two features directly connected by causal relationships increased category membership ratings above and beyond the presence (or absence) of each feature individually. In comparison, in the control condition the six contrasts were all approximately 0, indicating that correlations between features had no influence on category membership ratings in that condition.

A two-way ANOVA of these contrasts with condition (chain vs. control) and contrast (Δ_{21} , Δ_{32} , Δ_{43} , Δ_{31} , Δ_{42} , or Δ_{41}) as factors confirmed that the pattern of contrasts differed between the two conditions as indicated by a significant interaction between condition and contrast, $F(5, 350) = 4.53$, $MSE = 0.00551$, $p < .001$. In a separate analysis of the chain condition, the contrasts associated with directly connected feature pairs (Δ_{21} , Δ_{32} , Δ_{43}) were greater than those associated with indirectly connected pairs (Δ_{31} , Δ_{42} , Δ_{41}), $F(1, 35) = 12.0$, $MSE = 0.00551$, $p < .001$.

Theoretical modeling. To demonstrate that the pattern of results in Experiment 1 is in accordance with the predictions of causal-model theory, the chain model of Figure 1B was fit to the categorization ratings in the chain condition. That is, the values of parameters c , m , b , and K that minimize the squared difference

between the ratings and the model likelihoods as computed by the equations in Table 3 was determined for each of the 36 chain participants. The parameter values averaged over participants are presented in Table 5. The table also presents for each model the squared error averaged over participants (Avg. SSE), and the root-mean-square deviation (RMSD) averaged over participants (Avg. RMSD), where $RMSD = \sqrt{SSE/(N - P)}$, N is the number of observations modeled (16), and P is the total number of parameters in the model (4).

According to causal-model theory, parameter m represents the probability that a causal mechanism between two category features will operate (i.e., will bring about the presence of its effect) when the cause feature is present. Table 5 indicates an estimate of parameter m of .120. That is, the fit of the chain model indicates that participants generated categorization ratings in a manner consistent with a belief in the presence of probabilistic causal mechanisms arranged in a causal chain. To illustrate the chain model's sensitivity to correlations between features directly linked by causal mechanisms, its predicted ratings for exemplars 0000, 0101, 1010, and 1111 are presented in Figure 3, superimposed on the empirical ratings. The chain model correctly predicted both the lower categorization ratings for those exemplars that possessed many violations of causal knowledge (0101 and 1010) and the higher categorization ratings for those exemplars that confirmed causal knowledge (1111).

To demonstrate the chain model's fit directly in terms of feature probabilities and correlations, I computed the probabilities and correlations implied by the chain model fit. That is, for each participant the chain model predicts certain likelihoods for each exemplar, and feature probabilities and interfeature contrasts can be computed from these likelihoods. These probabilities and contrasts were then averaged over participants and these averages are presented in Figure 3B, superimposed on the empirical data. Figure 3B confirms that the chain model was able to account for both the weights associated with the individual features and the sensitivity to correlated features displayed by participants. Indeed, the chain model's predicted ratings accounted for 97% of the variance in the average ratings.

Table 5
Parameter Estimates and Measures of Fit of the Chain Causal Model to the Chain Conditions of Experiments 1–3

Parameters and measures	Experiment					
	1		2		3	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
<i>c</i>	.598	.017	.525	.016		
<i>m</i>	.120	.026	.402	.053	.266	.044
<i>b</i>	.521	.014	.384	.026	.464	.031
<i>K</i>	815	25.3	642	42.1	403	20.7
Average SSE	2157		2309		893	
Average RMSD	12.4		12.4		11.7	
<i>R</i> ²	.97		.93		.97	

Note. Average SSE = sum of squared error averaged over participants; Average RMSD = root-mean-square deviation averaged over participants. Parameter c was not estimated in Experiment 3.

Discussion

Experiment 1 confirmed causal-model theory's predictions regarding how causal knowledge induces both differences in feature importance and a sensitivity to correlated features. First, there was a significant difference in the weight of the first feature (F_1) compared with the remaining three features. Second, there were significant interactions between features directly connected by causal relationships. This overall pattern of results was observed in one of the examples presented in the Causal Models and Derived Empirical Statistics section of patterns that can be generated from a chain causal model: Case $c = .75$, $b = .50$, and $m = .10$ in Figure 2E.

Causal-model theory contrasts with alternative proposals that characterize the effects of knowledge solely in terms of changes to subjective feature weights (Keil, 1989; Medin & Shoben, 1988). For example, Ahn and her colleagues (Ahn, 1998; Ahn et al., 2000; Sloman et al., 1998) have proposed the causal status hypothesis, in which features are more important to categorization to the extent that they are "more causal," that is, more deeply embedded in a causal network as causes. For example, Ahn et al. (2000) found that when participants were taught novel categories with three features, X, Y, and Z, arranged in a causal chain (i.e., X causes Y, which causes Z), participants rated an exemplar missing Y as a worse category member than one missing Z, and an exemplar missing X worst of all, indicating that X was weighed more heavily than Y, which was weighed more heavily than Z.

The current results provide only partial support for the causal status hypothesis. First, whereas the feature that was the root cause in a causal chain (F_1) had greater weight on category membership ratings than the other three features, the more causal features (F_2 , F_3) were not weighed more heavily than the less causal feature (F_4). Second, the results of Experiment 1 indicate that participants also attended to whether pairs of features confirmed or violated the causal laws that linked features: Category membership ratings were higher to the extent feature pairs confirmed causal relationships (i.e., cause and effect features were both present or both absent) and were lower to the extent they violated those relationships (i.e., one feature was present and one was absent).

In contrast to the causal-status hypothesis, causal-model theory provides an integrated account of changes to feature weights and interfeature correlations brought about by causal knowledge. According to this account, people evaluate the probability that the entire combination of features displayed by an exemplar was likely to have been generated by the category's causal laws. As a result, causal-model theory was able to provide a good quantitative account of changes to both feature weights and correlations (see Figure 3B).

A basic claim of causal-model theory is that people view causal relationships as being constituted by probabilistic causal mechanisms, rather than by a relationship of necessity and sufficiency. The parameter estimates produced by the chain model fits provide information about how participants viewed the nature of the causal links. For example, a value of 1 for parameter m indicates that cause features are sufficient to bring about their effects. In fact, the average value of m was .120, and the largest value was .562. That is, participants assigned an exemplar some significant chance of category membership even if a cause feature was present and its effect was absent. For example, Exemplar 1110 received a fairly

high category membership rating of 71.4 even though it possessed one violation of sufficiency.

Similarly, a value of 0 for parameter b indicates that cause features are viewed as necessary for their effects to be brought about. In fact, the average value of b was .521, and the smallest value was .337. That is, participants assigned an exemplar some nonzero chance of category membership even if an effect was present and its cause absent. For example, Exemplar 0111 received a rating of 61.6 even though it possessed one violation of necessity. These results support the claim that people typically treat causal laws as probabilistic rather than as relationships of necessity and sufficiency.

Experiment 2

Experiment 2 replicated Experiment 1 in all respects except that the 75% base-rate information about features was omitted. There were two reasons for this manipulation. First, despite previous proposals suggesting that more causal features will be weighted more heavily in categorization judgments, in the Experiment 1 chain condition there was no difference in the weights associated with features F_2 , F_3 , and F_4 , even though F_2 was the cause of F_3 , which in turn was the cause of F_4 . One reason for this result may have been that participants were provided with information about not only interfeature causal links but also about the empirical base rates of features that indicated that all features were equally likely (i.e., 75%); this fact may have overwhelmed the efficacy of causal knowledge to elicit differences in feature weights for all but the most causal of features (F_1). This explanation was tested in Experiment 2 by omitting the base rates.

Second, as predicted by causal-model theory the causal knowledge in Experiment 1 induced a sensitivity to interfeature correlations for those feature pairs directly connected by causal relationships. Also as predicted by causal-model theory, the relatively weak strength of causal knowledge in Experiment 1 resulted in the absence of correlations between indirectly connected features. One reason for the weak strength of causal knowledge may have been that many chain participants attended only to the 75% base-rate information. Thus, the second purpose of omitting the feature bases was to increase the number of participants who used causal knowledge when categorizing, to determine whether the predicted correlations between indirectly connected features would be observed.

Method

Participants. Seventy-two University of Illinois at Urbana-Champaign undergraduates received course credit for their participation in this experiment. They were assigned in equal numbers to the chain and control conditions and to one of the six experimental categories.

Results

Category membership ratings for the 16 test exemplars averaged over participants in the chain and control conditions are presented in Table 4, and in Figure 4A for selected exemplars. Figure 4A demonstrates that the presence of causal knowledge linking category features had a large effect on the category-membership ratings participants gave the test exemplars relative to the control condition. As in Experiment 1, Exemplars 0101 and 1010 were

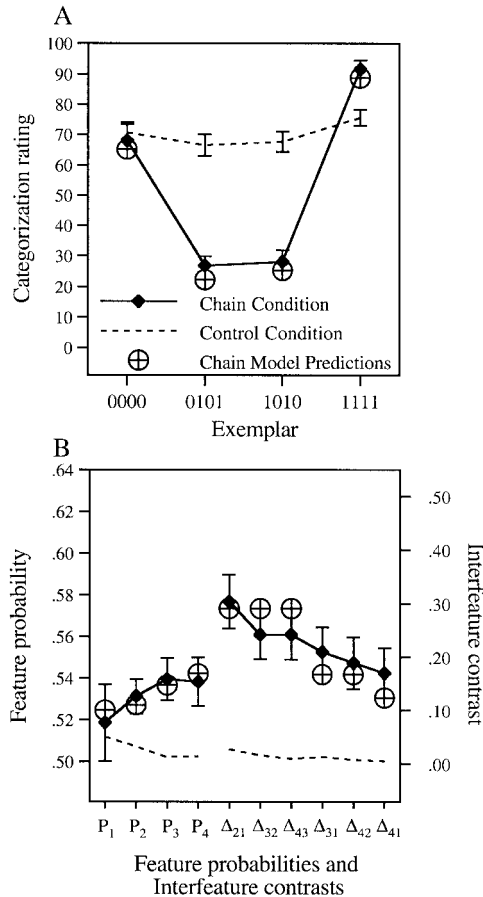


Figure 4. Results from the chain and control conditions of Experiment 2. A: Categorization ratings for selected exemplars. B: Derived feature probabilities and interfeature correlations. Predictions of the chain model are shown in each panel.

given much lower ratings in the chain condition (26.9 and 28.1, respectively) than in the control condition (70.7 and 70.6, respectively). In addition, exemplar 1111 received a higher rating in the chain versus control condition (91.6 vs. 76.4, respectively). A two-way ANOVA with condition (2 levels: chain vs. control) and test exemplar (16 levels) revealed a significant interaction, $F(15, 1050) = 17.28$, $MSE = 223.9$, $p < .0001$, confirming that the pattern of ratings given to the 16 test exemplars differed between the chain and control conditions.

Feature probabilities. As in Experiment 1, to evaluate the chain and control conditions in terms of their feature probabilities and interfeature correlations, each participant's set of categorization ratings was treated as a set of likelihoods by rescaling the ratings so they summed to 1. As in Experiment 1, feature probabilities and interfeature contrasts were then computed from those likelihoods. The four feature probabilities (P_1 , P_2 , P_3 , P_4) and the six interfeature contrasts (Δ_{21} , Δ_{32} , Δ_{43} , Δ_{31} , Δ_{42} , Δ_{41}) averaged over participants in the chain and control groups are displayed in Figure 4B.

One purpose of omitting the base-rate information was to determine whether the absence of empirical information about feature probabilities would result in features being weighed more

heavily to the extent that they were causes. In fact, Figure 4B indicates that more causal features tended to be weighed less heavily: The derived probabilities P_1 , P_2 , P_3 , and P_4 in the chain condition were .519, .531, .540, and .540, respectively. That is, not only were F_2 and F_3 not more heavily weighed than F_4 , the weight for the root cause feature F_1 was not greater than the weights of the other three features as it was in Experiment 1. As expected, given the absence of feature base-rate information, feature probabilities in the control condition were all approximately .50.

A two-way ANOVA of the feature probabilities with condition (chain vs. control) and feature (F_1 , F_2 , F_3 , and F_4) as factors was conducted. There was an effect of condition indicating that feature probabilities were greater in the chain ($M = .534$) versus the control condition ($M = .506$), $F(1, 70) = 12.6$, $MSE = 0.0038$, $p < .001$. However, despite the apparently increasing feature probabilities in the chain condition suggested by Figure 4B, the interaction between condition and feature did not approach significance, $F(3, 210) = 1.24$, $MSE = 0.0029$, $p > .20$.

Interfeature contrasts. The six interfeature contrasts (Δ_{21} , Δ_{32} , Δ_{43} , Δ_{31} , Δ_{42} , Δ_{41}) averaged over participants are also displayed in Figure 4B for the chain and control conditions. Figure 4B indicates that the pattern of these contrasts differed between the chain and control groups. In particular, in the chain condition the contrasts between those features directly connected in the causal chain (Δ_{21} , Δ_{32} , Δ_{43}) were .304, .242, and .242, respectively. These probabilities are each greater than 0, reflecting that participants produced higher categorization ratings when pairs of features were consistent with the causal laws and produced lower ratings when pairs were inconsistent with those laws.

The analyses of the chain model presented in the Causal Models and Derived Empirical Statistics section indicate that features indirectly connected by causal relationship can also be expected to be substantially correlated when causal mechanisms are sufficiently strong. Consistent with this claim, Figure 4B indicates that the contrasts between indirectly connected features (Δ_{31} , Δ_{42} , Δ_{41}) were each significantly greater than zero (.210, .189, and .189, respectively). As also predicted by the chain model, the contrasts between indirectly connected features ($M = .196$) were smaller than those between features directly connected ($M = .256$).

In contrast to the chain condition, in the control condition the six interfeature contrasts were all approximately 0, indicating that correlations between features had no influence on category-membership ratings in that condition. A two-way ANOVA of the interfeature contrasts with condition (chain vs. control) and contrast (Δ_{21} , Δ_{32} , Δ_{43} , Δ_{31} , Δ_{42} , or Δ_{41}) as factors confirmed that the contrasts were larger in the chain condition than in the control condition, $F(1, 70) = 21.48$, $MSE = 0.226$, $p < .0001$. Further, a significant interaction between condition and contrast, $F(5, 350) = 4.29$, $MSE = 0.00528$, $p < .005$, indicated that the pattern of contrasts differed between the two conditions. In particular, in a separate analysis of the chain condition the contrasts associated with directly connected feature pairs (Δ_{21} , Δ_{32} , Δ_{43}) were significantly greater than those associated with indirectly connected pairs (Δ_{31} , Δ_{42} , Δ_{41}), $F(1, 35) = 19.14$, $MSE = 0.00335$, $p < .0001$.

Theoretical modeling. As in Experiment 1, the chain model was fit to the categorization ratings of each participant in the chain condition. The parameter values are presented in Table 5, and the fit of the chain model is illustrated in Figure 4A for selected

exemplars. As in Experiment 1, the estimate for parameter m was .402, indicating that participants' categorization ratings were consistent with the presence of probabilistic causal mechanisms. In particular, because of its sensitivity to correlated features, the chain model accounts for those exemplars that possess many confirmations of causal relationships (0000 and 1111) as well as for those that possess many violations of causal relationships (0101 and 1010).

As in Experiment 1, the feature probabilities and interfeature correlations implied by the model fits were computed for each participant and then averaged over participants; they are presented in Figure 4B, superimposed on the empirical data. Figure 4B confirms that the chain model provided a reasonable account for the weights associated with the individual features displayed by participants. In addition, both the chain model's predictions and the observed data exhibit strong correlations among features that are directly connected by causal relationships, and weaker correlations among features that are indirectly connected. As a result, the chain model's predicted ratings accounted for 93% of the variance in the average ratings in the chain condition. One potential discrepancy between the observed and predicted ratings is that the chain model tends to overestimate directly connected features contrasts and to underestimate indirectly connected contrasts.

Discussion

One purpose of Experiment 2 was to assess the participants' use of causal laws in the absence of explicit base-rate information about features. The chain participants in Experiment 2, as compared with those in Experiment 1, made more use of causal laws when rating category membership. This result was reflected in the larger estimate for the chain model's m parameter in Experiment 2 (.402) as compared with that in Experiment 1 (.120). The greater use of causal knowledge manifested itself in the predicted pattern of interfeature correlations, that is, there were large correlations between feature pairs directly connected by causal mechanisms, and there were weaker (but still substantial) correlations between pairs that were indirectly connected.

Another purpose of omitting feature base-rate information in Experiment 2 was to determine whether intermediate cause features (F_2, F_3) would become more heavily weighed than the effect feature (F_4). In Experiment 2, all four features were weighed approximately equally, whereas in Experiment 1, feature F_1 was weighed more heavily than the other three features.

As in Experiment 1, causal-model theory provided a good quantitative account of the Experiment 2 category-membership ratings. In particular, the theoretical entities postulated by causal-model theory, such as causal mechanisms, were able to account for both the pattern of feature weights and the interfeature correlations found in Experiment 2. Indeed, the pattern of results in Experiment 2 corresponded to one of the chain model examples presented in the Causal Models and Derived Empirical Statistics section: Case $c = .50, b = .50, \text{ and } m = .40$ in Figure 2F.

Implicit Causal Models

One result obtained from Experiment 2's chain condition was that feature weights did not vary as a function of position in the causal chain. This result is surprising in light of past studies that

have found that features that are more causal (more deeply embedded in a causal chain) are weighed more heavily (Ahn, 1998; Ahn et al., 2000; Sloman et al., 1998; present Experiment 1). Therefore, it is important to consider reasons that more causal features are often weighed more heavily in categorization decisions and why those reasons may not have applied in Experiment 2. One explanation of why causal features receive more weight than other features is that participants often reason with a more complex causal model than the one with which they were explicitly provided. For example, one of the novel categories used in Ahn et al. (2000) was a novel disease, D , which was described as causing a symptom, X , which caused another symptom, Y , which caused a third symptom, Z . (Realistic-sounding medical terms, rather than $D, X, Y, \text{ and } Z$, were used for the disease and symptoms.) Ahn et al. assumed that participants' causal knowledge consisted of $X \rightarrow Y \rightarrow Z$. However, people's default understanding of the relationship between diseases and symptoms is that a disease causes its symptoms, so participants were likely to have induced the causal model $D \rightarrow X \rightarrow Y \rightarrow Z$ where D (the disease) is the defining feature of the category (e.g., a virus). Under these circumstances participants may have reasoned backward from the symptoms to the disease, and, of course, symptoms closer (in a causal sense) to the disease were taken to be more diagnostic of that disease. Several examples of such decreasing feature weights are presented in Figures 2A and 2D. In these panels, the disease is a root cause that is invariably present ($c = 1$).

Participants may have also assumed more complex causal models for the novel natural kinds and artifacts used in Ahn (1998), Ahn et al. (2000), and the current Experiment 1. Researchers have suggested that people view many kinds as being defined by underlying properties or characteristics (sometimes referred to as an *essence*) that are shared by only category members and by members of no other categories (Gelman & Wellman, 1991; Keil, 1989; McNamara & Sternberg, 1983; Rips, 1989). Moreover, essential features are presumed to generate, or cause, perceptual features (Gelman, Coley, & Gottfried, 1994; Medin & Ortony, 1989; Rehder & Hastie, 2001). Although many artifacts such as pencils and wastebaskets do not appear to have internal causal mechanisms, it has been suggested that the essential properties of such artifacts may be the causal force responsible for their existence, namely, the intentions of the artifact's designer (Bloom, 1998; Keil, 1995; Matan & Carey, 2001; Rips, 1989). In each of these cases, the causal model that people use during categorization may have also included the background causes that categorizers presume bring rise to a category's observable features.

One reason that participants in Experiment 2 may not have weighed causes more heavily is that they may have been less likely to consider background causes that define category membership. This may have occurred because they may have assumed that the so-called features of the category were in fact rare exceptions. For example, participants were told that some Lake Victoria Shrimp have high amounts of ACh neurotransmitter and that some have normal amounts. Under these conditions, some participants might have thought that having high amounts of ACh was exceptional and hence not produced by the normal causes of features of biological kinds such as DNA. Indeed, in support of this interpretation a substantial minority of the chain participants in Experiment 2 ($n = 5$) assigned very low weights to the first feature ($F_1 < .40$). As a result of considering features to be exceptional rather

than characteristic, these 5 participants may not have reasoned with a causal model that included an invariant cause of those features, as did many participants in Experiment 1 and previous studies.

In Experiment 3, I test the hypothesis that when a chain of category features is assumed to be produced by an underlying cause, the result is the pattern of decreasing feature weights that has been taken as evidence for the causal-status hypothesis. This test was accomplished by describing each category as possessing an essential feature that is the cause of the other features.

Experiment 3

The materials from Experiments 1 and 2 were altered so that F_1 was the defining feature of the category. This was accomplished by stating that F_1 occurred in 100% of all category members and that it did not occur in members of other categories and by changing the name of the category. For example, *Lake Victoria Shrimp* were renamed to *Acetylcholine Shrimp* and participants were told that all acetylcholine shrimp have high amounts of acetylcholine and that no other kind of shrimp does. Note that the resulting categories correspond to the example introduced in the beginning of this article: Bird DNA causes wings, which cause flying, which causes nests in trees (i.e., bird DNA plays the role of the defining characteristic of birds). Because in the real world, people usually cannot directly observe essential features (such as DNA), the test exemplars in Experiment 3 listed the value on the first dimension as unknown.

Method

Materials. The categories were renamed as follows: *Kehoe Ants*, *Lake Victoria Shrimp*, *Myastars*, *Meteoric Sodium Carbonate*, *Romanian Rogos*, *Neptune Military Personal Computers* became *Iron Sulfate Ants*, *Acetylcholine Shrimp*, *Ionized Helium Stars*, *Radioactive Sodium Carbonate*, *Butanos*, and *Magnetic Computers*. For each category, F_1 was described as occurring 100% of the time in category members and never in members of other categories. As in Experiment 2, no base-rate information was provided for features F_2 , F_3 , or F_4 . As in Experiments 1 and 2, participants in the chain condition were also told of causal relationships between F_1 and F_2 , F_2 and F_3 , and F_3 and F_4 .

Procedure. The training procedure was identical to that in Experiments 1 and 2. After training, participants were presented with test exemplars in which the presence or absence of the first feature was always unknown (as designated by ???). That is, participants rated the category membership of eight three-feature exemplars, x000, x001, x010, x100, x011, x101, x110, and x111, and six single-feature exemplars, x0xx, x1xx, xx0x, xx1x, xxx0, and xxx1 (in which x denotes an unknown value), each presented twice. As in Experiments 1 and 2, presentation of the ratings for the single-feature exemplars is omitted.

Participants. Fifty-four University of Illinois at Urbana-Champaign undergraduates received course credit for their participation, with participants randomly assigned to the chain ($n = 36$) or control ($n = 18$) conditions. In equal numbers, participants were randomly assigned to one of the six categories.

Results

Category-membership ratings for the eight three-feature test exemplars averaged over participants in the chain and control conditions are presented in Table 6, and in Figure 5A for selected exemplars. Figure 5A indicates that in Experiment 3, as in Exper-

Table 6
Exemplar Classification Ratings From Experiment 3

Exemplar	Chain		Control	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
x000	45.8	4.3	57.6	4.4
x001	37.4	3.4	63.5	4.3
x010	35.8	3.2	63.6	4.0
x100	43.9	3.2	64.3	4.4
x011	44.4	4.2	67.9	3.6
x101	48.9	4.4	66.1	4.6
x110	54.6	4.1	68.1	4.0
x111	93.8	1.7	72.5	4.2

iments 1 and 2, the presence of causal knowledge linking category features affected the category membership ratings participants gave the test exemplars. Exemplars that violated causal relationships (e.g., x010 and x101) were given lower ratings in the chain condition (35.8 and 48.9, respectively) than in the control condition (63.6 and 66.1, respectively). In addition, a test exemplar that confirmed many causal relationships (x111) was given a higher rating in the chain condition (93.8) than in the control condition (72.5). A two-way ANOVA with Condition (chain vs. control) and Test Exemplar (8 levels), revealed a significant interaction, $F(7, 364) = 13.37$, $MSE = 220.90$, $p < .0001$, confirming that the pattern of ratings differed between the chain and control conditions.

Feature probabilities. As in Experiments 1 and 2, each participant's set of category-membership ratings was treated as a set of likelihoods by rescaling the ratings so that they summed to 1. Because test exemplars did not vary the presence or absence of F_1 it was not possible to compute P_1 or the interfeature contrasts involving F_1 . The three remaining feature probabilities (P_2, P_3, P_4) and the three remaining contrasts ($\Delta_{32}, \Delta_{43}, \Delta_{42}$) averaged over participants in the chain and control groups are displayed in Figure 5B.

The purpose of Experiment 3 was to determine whether an unobserved essential feature would generate the expected pattern of decreasing feature weights across the causal chain. Figure 5B confirms this prediction, because F_2 was more probable (.611) than F_3 (.577), which was more probable than F_4 (.563). In contrast, in the control condition, the weights associated with features did not differ appreciably from one another, .517, .522, and .517 for $F_2, F_3,$ and F_4 , respectively. A two-way factorial ANOVA of the feature probabilities, with condition (chain vs. control) and feature (F_2, F_3, F_4) as factors, revealed that feature weights were significantly larger in the chain condition than in the control condition, $F(1, 52) = 10.34$, $MSE = 0.0147$, $p < .005$. There was also an interaction between condition and feature, $F(2, 104) = 4.38$, $MSE = 0.00175$, $p < .05$, indicating that the pattern of feature probabilities differed in the two conditions. A separate analysis of the chain condition confirmed that the probability of F_2 was significantly greater than the average probability of F_3 and F_4 , $F(1, 35) = 11.65$, $MSE = 0.00520$, $p < .005$. The difference between F_3 and F_4 was marginally significant, $F(1, 35) = 2.61$, $MSE = 0.00247$, $p < .10$.

Interfeature contrasts. Figure 5B also indicates that the pattern of the three interfeature contrasts $\Delta_{32}, \Delta_{43}, \Delta_{42}$ differed between

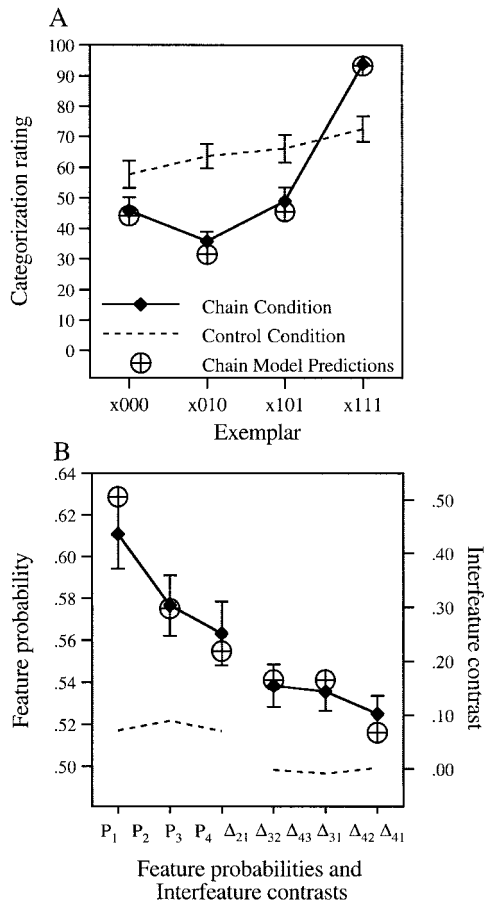


Figure 5. Results from the chain and control conditions of Experiment 3. A: Categorization ratings for selected exemplars. B: Derived feature probabilities and interfeature correlations. Predictions of the chain model are shown in each panel.

the chain and control groups. Moreover, in the chain condition the contrasts between those features directly connected in the causal chain (Δ_{32} , Δ_{43}) were .154 and .144, respectively, and the contrast between the indirectly connected features (F_2 and F_4 ; Δ_{42}) was smaller (.102) than those between the directly connected features. In the control condition the three interfeature contrasts were all approximately zero, indicating that correlations between features had no influence on category-membership ratings in that condition. A two-way factorial ANOVA of the interfeature contrasts, with condition (chain vs. control) and contrast (Δ_{32} , Δ_{43} , Δ_{42}) as factors, confirmed that the contrasts were larger in the chain versus the control condition, $F(1, 51) = 8.49$, $MSE = 0.086$, $p < .01$. A separate analysis of the chain condition indicated that the difference between contrasts associated with directly connected-feature pairs (Δ_{32} and Δ_{43}) and with the indirectly connected pair (Δ_{42}) was marginally significant, $F(1, 34) = 2.82$, $MSE = 0.00441$, $p < .15$.

Theoretical modeling. The likelihood equations for the chain model for the eight test exemplars with an unknown value on the first dimension were derived from those for the four-feature exemplars in Table 3 by summing across the two possible values of

the unknown feature. For example, $L(x000) = L(0000) + L(1000)$. However, because F_1 was described as occurring 100% of the time, I assume that Parameter $c = 1$, and hence $L(0000) = 0$ and $L(x000) = L(1000)$. Note that this model has one fewer free parameter than the first two experiments because c is assumed to equal 1. The resulting parameter estimates are presented in Table 5.

As in Experiments 1 and 2, the estimate for parameter m (.266) indicated that participants' categorization ratings were consistent with the presence of probabilistic causal mechanisms. Figure 5A indicates that, because of its sensitivity to correlated features, the chain model was able to account for the category-membership ratings of those exemplars that confirm causal laws (x000 and x111) as well as for those that violate those laws (x010 and x101).

The feature probabilities and interfeature contrasts implied by the chain model fits were also computed for each participant, and the probabilities and contrasts averaged over participants are presented in Figure 5B. Figure 5B indicates that the chain model reproduces the pattern of decreasing feature weights seen in the empirical data. It also indicates that, as in Experiments 1 and 2, the chain model accounts for participants' sensitivity to correlated features. Although the chain model tends to underestimate indirectly connected contrasts (as it did in Experiment 2), it nevertheless accounted for 97% of the variance in the average ratings.

Discussion

The goal of Experiment 3 was to induce the expected pattern of feature weights when the root of a causal chain is the essential feature. As predicted, the features in the causal chain (F_2 , F_3 , F_4) were more heavily weighted than in the control condition, and those weights decreased as a feature became more distant from the defining feature (F_1). Several examples of such decreasing feature weights with a chain model were presented earlier in Figures 2A and 2D.

Previous research (Ahn, 1998; Ahn et al., 2000, and the present Experiment 1) has found that features in the early part of a causal chain are weighed more heavily than other features. The results of Experiment 3 suggest that this effect arises when categorizers reason with a causal model with a hidden and invariant root cause, such as when an unobservable disease, D , causes a symptom, X , which causes another symptom, Y , which causes yet another symptom, Z , or when an unobservable but causally potent essential feature produces a chain of features. Causal-model theory formalizes the bit of common-sense causal reasoning that tells us that symptoms are produced more reliably to the extent that they are causally closer to the disease, or that object features are produced more reliably when they are causally closer to a kind's causally potent defining feature. In contrast, many participants in Experiment 2 may not have considered the implicit background causes of category features because the features were considered to be exceptional rather than characteristic of the category.

As in Experiments 1 and 2, in Experiment 3 the interfeature contrasts between feature pairs directly connected by causal relationships were significantly greater than zero. This result obtained despite the presence in Experiment 3 of a defining feature (F_1), which one might have predicted would make evidence of the operation of a remote causal mechanism (e.g., between F_3 and F_4) irrelevant. In other words, participants did not just use causal

relationships to infer the presence of an invisible defining feature, but evidence of the presence of the causal mechanisms themselves also made exemplars better category members. In addition, as in Experiments 1 and 2, the interfeature contrast between the indirectly connected features F_2 and F_4 was less than the contrast between directly connected features (although the difference was only marginally significant in this experiment).

Evaluating a Similarity-Based Account of Correlated Features

One of the primary results of the current experiments was that participants' category-membership ratings were sensitive to correlated features that either confirmed or violated experimentally provided causal laws. Although similarity-based exemplar models (Medin & Schaffer, 1978; Nosofsky, 1986) are well known for their ability to exhibit sensitivity to correlated features as a by-product of computing similarity to category exemplars stored in memory, in the current experiments participants observed no category exemplars. Thus, the effects found in the chain condition of Experiments 1–3 must be attributed to the causal knowledge itself rather than to previously observed category members.

The inability of similarity-based prototype and exemplars models to account for correlated features would seem to undermine the role of similarity in accounting for categorization decisions in the presence of causal knowledge about a category. Nevertheless, similarity-based models grant themselves considerable flexibility by leaving open the question of what counts as a feature. For example, to account for the effects of interfeature correlations induced by causal knowledge within a similarity-based framework, one might postulate the existence of second-order features (Gluck & Bower, 1988; Hayes-Roth & Hayes-Roth, 1977; Minsky & Papert, 1988; Neumann, 1974; Reitman & Bower, 1973; Rumelhart, Hinton, & Williams, 1986) that (a) encode the confirmation or violation of causal relationships and (b) participate in similarity computations alongside primitive features. According to this account, the encoding of, say, exemplar 0110 in the chain conditions of Experiments 1–3 would have included second-order properties indicating that the first and third causal links were violated (F_1 absent but F_2 present, F_3 present but F_4 absent) and that the second link was confirmed (F_2 and F_3 both present). Categorization ratings were then computed on the basis of the similarity between the exemplar (0110010) (with the last three features 010 indicating which causal links are confirmed or violated), and a seven-dimension category prototype (1111111) that represents the typical category member in which all four features are present and all three causal relationships are confirmed.

This proposal, which hereinafter I refer to as the *configural-features prototype model*, is formalized in the Appendix. As reported there, the configural-features prototype model was fit to the category membership ratings of the 32 participants in the chain condition of Experiment 3. The results indicated that the configural-features prototype model yielded fits approximately as good as the chain model in Experiment 3. In particular, it accounted for the categorization ratings of exemplars x000, x010, x101, and x111 presented in Figure 5A.

The difference is that although the configural-features prototype model accounts for feature weights and interfeature correlations, causal-model theory *explains them* in terms of a deeper theoretical

construct, a causal model. That is, whereas the configural-features prototype model achieved its fit to the Experiment 3 chain-rating data at the cost of extra parameters that represent the weight of each feature, the chain model reproduces both the pattern of feature importance and interfeature correlations because they are an emergent property of causal models rather than results of either explicit feature weights or second-order properties. As a result, the chain model has fewer parameters than the configural-features prototype model and thus achieves a better average RMSD and a better fit (according to RMSD) for 27 of the 36 chain participants in Experiment 3.

General Discussion

These results support the claim that people classify new objects by evaluating whether the combination of features displayed by an object was likely to have been generated by a category's causal laws. In three experiments, causal knowledge provided as part of the experimental session was found to have large effects on subsequent category-membership ratings, effects which were qualitatively consistent with those expected when features are arranged in a causal chain. Moreover, causal-model theory provided a good quantitative fit to the categorization ratings from all three experiments.

Causal-model theory's good fits were reflected in its ability to successfully model the changes brought about by causal knowledge to both the individual feature weights and the interfeature correlations observed in participants' categorization ratings. However, those fits were not obtained because causal-model theory possesses free parameters that correspond to feature weights and correlations. Instead, the claim is that categorizers judged whether an exemplar was generated by causal laws, and a sensitivity to feature weights and the interfeature correlations emerged as a side effect of that process; thus, causal-model theory explains the sensitivity to features and correlations implicit in categorization judgments in terms of their classifiers of a deeper theoretical construct, namely, a causal model.

In the following subsection, I discuss the effect of causal knowledge on feature weights and interfeature correlations. I then discuss the ability of causal-model theory to represent empirical in addition to causal information. Finally, I discuss causal-model theory's representation of causal knowledge.

Causal Knowledge and Feature Weights

One important empirical finding from recent years has been the demonstration that the underlying attributes that objects are thought to possess are critical for establishing the object's category membership. In a well-known study, Rips (1989) found that undergraduates judged that animals that undergo dramatic changes of appearance still retain their category membership (e.g., a bird that takes on insect features because of exposure to chemical waste is still judged to be a bird; also see Gelman & Wellman, 1991; Keil, 1989). Moreover, Medin and Ortony (1989) have argued that underlying properties not only establish category membership, they "are best thought of as constraining or even generating properties that might turn out to be useful in identification." Further, features "[lie] on a continuum of centrality ranging from

relatively inaccessible, deep properties to more accessible, surface ones" (p. 185).

Causal-model theory formalizes the relationship between observable features, on the one hand, and the central but unobservable features that generate them, on the other. By assuming that a feature's centrality is determined by the likelihood that it was generated by a category's causal mechanisms, causal-model theory explains why features should be considered central to the extent they are causally closer to an underlying mechanism that defines category membership. In fact, I found in Experiment 3 the pattern of decreasing feature centrality predicted by causal-model theory when the chain categories were described as possessing an invariant root cause. Causal-model theory predicts this effect because it incorporates the causal reasoning processes that tell us, for example, that symptoms that are more directly caused by a disease are more reliable predictors of that disease.

By specifying that a feature's centrality is determined by underlying causal mechanisms, causal-model theory differs from alternative proposals like the causal-status hypothesis that suggests that features will always be more heavily weighed to the extent they are more deeply embedded as causes in a causal network (Ahn, 1998; Ahn et al., 2000; also see Sloman et al., 1998). Indeed, the present results do not provide consistent support for the causal-status hypothesis, because although the root cause feature (F_1) was weighed the most heavily in Experiment 1, the intermediate cause features (F_2 , F_3) were not more heavily weighed than F_4 . In addition, the causal-status hypothesis was absent entirely in Experiment 2. Causal-model theory attributes this weakened or absent causal-status effect to participants' not linking the root of the causal chain to the category's underlying causal mechanisms. When such linkage was made explicit (as in Experiment 3), the causal-status effect found in previous studies emerged.

The importance of a category's causal laws in determining whether a feature is likely to be present has been demonstrated in other studies. For example, using the same materials as in the present experiments, Rehder and Hastie (2001) and Rehder (2003) taught participants a *common-cause schema* in which one category feature directly causes all other features, and a *common-effect schema* in which one category feature is caused by all other features. Whereas the results from the common-cause schema supported the causal-status hypothesis (i.e., the common-cause feature was weighed most heavily), the common-effect feature was found to be weighed much more heavily than its causes in the common-effect schema, a result causal-model theory attributes to categorizers reasoning that the common-effect feature will be present in many category members because it will be frequently produced by at least one of its multiple causes. Similarly, Rehder and Burnett (2002) found that participants were much more likely to infer the presence of an unobserved feature when its cause was present. The importance of causal knowledge also applies when projecting novel properties to other category members. For example, Lassaline (1996) found that participants were more likely to judge that an object possessed a novel feature when the purported cause of the feature was also present in the object. Research has generally shown that people will license those category-based property generalizations that are supported by the causal explanations their domain knowledge provides (Lopez, Atran, Coley, Medin, & Smith, 1997; Proffitt, Coley, & Medin, 2000).

By assuming that the presence of a feature is determined by the underlying causes that generate it, causal-model theory differs from similarity-based models that account for differences in feature centrality by postulating the presence of explicit weight parameters for each feature (Medin & Schaffer, 1978; Nosofsky, 1986). Because it explains differences in feature centrality in terms of causal mechanisms rather than explicit feature weights, causal-model theory provides a more parsimonious account (i.e., one that requires fewer parameters) of Experiment 3's categorization ratings as compared with a similarity-based model using per feature weights.

Causal Knowledge and Feature Configurations

Besides its novel predictions regarding the importance of individual features, causal-model theory also accounts for the effects of causal knowledge on specific configurations of features. According to this account, category membership is not just a matter of observing a category's most probable features, but also the most probable configuration of those features. Consistent with this prediction, in three experiments participants' category-membership ratings were sensitive to the correlations between features directly connected by causal relationships. In Experiments 2 and 3, participants' category-membership ratings were also sensitive to the weaker correlations expected to hold between indirectly connected features.

The use of a defining feature (F_1) in Experiment 3 presented a strong test of the claim that people categorize on the basis of whether the entire configuration of features displayed by an exemplar makes sense in light of a category's causal knowledge. If participants were only concerned with the presence or absence of the defining feature (F_1) then all the weight should have been assigned to F_2 , because only F_2 has any direct bearing on the presence of F_1 . In fact, participants were also sensitive to the correlations between F_2 and F_3 , and between F_3 and F_4 , because they increased (decreased) categorization ratings when those correlations were preserved (broken). That is, categorization judgments were based on not only evidence related to the presence of the unobserved defining feature (F_1), but also on evidence related to the presence of unobserved causal mechanisms. Apparently, evidence for a category's causal mechanisms in addition to any (so-called) defining features is required to make the best category members. Rehder and Hastie (in press) referred to such exemplars as *theoretical ideals* because they instantiate the full set of theoretical knowledge one should possess about a category.

An important factor controlling causal-model theory's predictions regarding the expected pattern of correlations among features is the asymmetry of causal knowledge, a characteristic that distinguishes causal-model theory from other models that represent theoretical relationships symmetrically (e.g., Heit, 1994; Rehder & Murphy, in press; and the configural-features prototype model presented earlier). The Rehder (in press) study demonstrated that representing causality as an asymmetric relation is necessary for explaining participants' sensitivity to configurations of features, because in that study, participants exhibited sensitivity not only to the expected pattern of pairwise correlations for both common-cause and common-effect schemas but also to higher order interactions among features for the common-effect network (i.e., interactions that correspond to the discounting effect frequently

observed in cases of multiple causation; Morris & Larrick, 1995). Such higher order interactions did not obtain for the common-cause schema even though it is the mirror image of the common-effect schema if one ignores the direction of causality. Models that ignore the direction of causality are unable to account for differences between common-cause and common-effect networks. In contrast, causal-model theory predicts both the observed higher order feature interactions found for common-effect networks and their absence from common-cause networks (Rehder, 2003).

Empirical Observations and Causal Models

Because the primary goal of this research is to establish the independent influence of causal knowledge on categorization, the current experiments differed from many past studies by not presenting participants with examples of category members. On the one hand, studying the effects of theoretical knowledge in isolation is interesting in its own right, as there are many real-world categories about which people know far more than they have observed first hand, such as scientific concepts (subatomic particles, galaxies, viruses), unusual psychological syndromes (schizophrenia, multiple personalities), types of rare political and economic events (revolutions, depressions), and national, ethnic, and racial stereotypes of remote peoples (for most Americans: Buddhists, North Koreans), to name a few. Causal-model theory accounts for such categorization decisions even when no examples of the category have been observed, a domain beyond the purview of traditional models that are solely based on empirical information.

On the other hand, for many other categories people observe many examples of category members, and numerous studies have demonstrated that that empirical information exerts an influence on categorization even when theoretical or causal knowledge about the category is also present (Heit, 1994; Kaplan & Murphy, 2000; Rehder & Hastie, 2001; Spalding & Murphy, 1999; Wisniewski, 1995; Wisniewski & Medin, 1994). Although in the current study, the parameters of causal models were intended to reflect only participants' causal knowledge, those parameters could be updated to reflect the empirical regularities that categorizers observe. For example, the c and b parameters of a causal model are intended to represent the probability that some unspecified background cause brings about a feature. However, categorizers' expectations regarding the strength of background causes are often based on the frequency with which features are observed. Indeed, in the current experiments the average values of the c and b parameters were greater when participants were told that features occurred with probability 75% in Experiment 1 (0.598 and 0.521, respectively) than when they were provided with no base-rate information in Experiment 2 (0.525 and 0.384). More generally, if participants had been presented with examples of category members, the chain model (see Figure 1B) could have been generalized to accommodate features that varied in their empirical frequency by allowing the b parameters associated with the three effects to vary independently (so that each feature had its own parameter: c for F_1 , and one b each for F_2 , F_3 , and F_4).

Similarly, the m parameters of causal models are intended to represent the probability that a causal mechanism brings about its effects. However, categorizers' beliefs regarding the strength of causal mechanisms may reflect the co-occurrence of cause and effect features that appear in observed category members. Al-

though the chain model accounted quantitatively for the results from Experiments 1–3, assuming that causal links were of equal strength (i.e., had the same m parameter), this success was likely due to the use of novel causal mechanisms that were pretested to be of equal plausibility. To accommodate causal mechanisms of unequal strength (due, for example, to the observation of interfeature correlations of unequal magnitude), the chain model could be generalized to allow the m parameter associated with each causal link to vary independently. Indeed, current theories of how causal strengths are estimated on the basis of empirical information (discussed in the next section) provide one mechanism by which the strengths of causal mechanisms might be updated on the basis of empirical evidence.¹ Thus, with time, categorizers can gradually refine the parameters of their initial causal model (the cs , ms , and bs) of a category to better reflect the empirical regularities they observe.

Representing Causal Knowledge

The use of causal models has been implicated in a number of other cognitive tasks. For example, Glymour and Cheng (Glymour, 1998; Glymour & Cheng, 1998) have shown that Cheng's (1997) causal power theory of causal induction can be derived from assumptions regarding the causal model people possess. They have argued that people's inductive reasoning is based on a *noisy or-gate* (Pearl, 1988), in which the effect may be caused not only by the cause of interest but also by some unspecified background causes. Cheng (1997) has derived how the strength of a causal mechanism that links a cause and effect is related to a measure of correlation between those variables (the probabilistic contrast).

The formalization of causal models offered by Glymour and Cheng (Glymour, 1998; Glymour & Cheng, 1998) is equivalent to the development presented here. For example, the simple causal model of Figure 1A is also a noisy or-gate in which Feature C causes Feature E, but Feature E might also be brought about by some unspecified background cause. Further, the relationship between the probabilistic contrast between C and E and the strength of the causal mechanism connecting them can be derived from Equation 4 and the exemplar likelihoods for Figure 1A's simple model presented in Table 2: $\Delta_{CE} = P(E|C) - P(E|\sim C)$ and $P(E|C) = P(11)/[P(11) + P(10)]$ and $P(E|\sim C) = P(01)/[P(01) +$

¹ Although it has been shown that categorizers are sensitive to empirical feature frequencies in the presence of causal knowledge (Spalding & Murphy, 1999; Wisniewski, 1995; Rehder & Hastie, 2001), there is reason to doubt they are sensitive to empirical feature correlations in a manner that can lead them to update their estimates of causal strengths. For example, Rehder and Hastie (2001) systematically examined the effects of causal knowledge and empirical observations and found that the presence or absence of interfeature correlations in observed category members had little effect on subsequent classification performance when causal knowledge was present. Similarly, Murphy and Wisniewski (1989) found that participants were sensitive to those interfeature correlations that were expected on the basis of prior knowledge but insensitive to those in empirical data (also see Chapman & Chapman, 1967, 1969). Thus, although it is well-established that categorizers exhibit sensitivity to interfeature correlations by similarity matching with category exemplars stored in memory (Medin & Schaffer, 1978; Nosofsky, 1986), whether causal strengths (i.e., m parameters) get updated on the basis of such correlations awaits explicit experimental demonstration.

$P(00)$]. Substituting in the expressions for $P(00)$, $P(01)$, $P(10)$ and $P(11)$ from Table 2 and simplifying yields

$$\Delta_{CE} = m(1 - b),$$

and expressing m as a function of Δ_{CE} and noting that $b = P(E|\sim C)$ gives

$$m = \Delta_{CE}/[1 - P(E|\sim C)],$$

which is Cheng's (1997) Equation 8 with p_i renamed m . That is, the probabilistic contrast measure used here and in Cheng's equation is taken as reflecting the operation of the same underlying theoretical construct: the power of a probabilistic causal mechanism.

The parameter estimates derived in the current study provide support for the view that people often see causal relationships as being constituted by probabilistic causal mechanisms, rather than as a relation of necessity and sufficiency. First, in all three experiments, the estimates for parameter m were less than 1, indicating that virtually all participants treated the interfeature causal relationships in terms of causal mechanisms that do not operate invariantly. This result contrasts with the frequent treatment of causal knowledge in terms of causes that deterministically yield their effects (e.g., Mackie, 1974). Second, the estimates for parameter b were greater than 0, indicating that almost all participants allowed for the possibility that an effect might have been brought about by some cause not explicitly specified as part of the category's causal knowledge.

Whereas the representation of causal knowledge proposed here is concerned with binary features or events, many real-world categories include other types of features and causal mechanisms involving more than just pairs of features. Fortunately, it is straightforward to represent these more complex structures in the Bayesian network framework adopted by causal-model theory. For example, rather than binary variables that represent the presence or absence of features, causal models can include binary variables with high–low semantics (in which, e.g., the high value of the cause variable produces one effect and the low value produces another) and variables that are ordinal, nominal, or continuous. When the causal influence between continuous variables is linear, then causal models correspond to a form of structural equation models. Indeed, using categories with continuously valued dimensions, Waldmann et al. (1995) induced participants to believe that a category possessed either a common-cause or a common-effect model, and found that participants' category-learning rates depended on whether exemplars manifested the correlational structure that those causal models led them to expect. Causal mechanisms may also involve more than two variables, as with a conjunctive cause in which each cause (e.g., fuel, oxygen, spark) is a necessary precondition for the effect (e.g., fire).

Causal-model theory can provide predictions for how people classify objects using networks with (a) ordinal, nominal, or continuous variables and (b) causal mechanisms involving more than two variables, because likelihood equations may be derived that represent the probability that a set of variable states will occur. Moreover, because those networks are based on the *structure* of causal knowledge (i.e., the strength and topology of the causal mechanisms) while abstracting over its content (e.g., the details of the causal mechanisms involved), causal-model theory is applica-

ble to any domain in which people have beliefs about the operation of causal mechanisms. Indeed, evidence in favor of this domain-general account is provided by the fact that the effects of causal knowledge in the present study were consistent across novel categories from a number of different domains, including biological kinds, nonliving natural kinds, and artifacts.

Finally, it may not be the case that all the "theoretical" knowledge one might ascribe to people is necessarily causal in nature. For example, besides causal–mechanical explanations, human knowledge includes ontological beliefs about the types of entities in the world (Chi, 1993; Keil, 1979) and teleological beliefs about the purposes that objects and their properties serve (Keil, 1995). The need to distinguish between causal–mechanical understandings (a *naive physics*) and those involving intentional agents (a *naive psychology*) has also been stressed (Carey, 1985, 1995). Nevertheless, a wide range of theorists have emphasized the special importance of causal relations in people's intuitive theories of the world (Carey, 1995; Gelman & Kalish, 1993; Gopnik & Wellman, 1994; Keil, 1989; Murphy & Medin, 1985). The special status granted causal knowledge should be unsurprising in light of its distinct, functional advantages. Although after-the-fact explanation may be satisfying (Gopnik, 2000), it is an ability to represent causal regularities that enables an organism to successfully intervene in external events and attain control over its environment (Sperber, Premack, & Premack, 1995).

Conclusion

The current article proposes a representation of the causal knowledge that interrelates or links the features of many categories and demonstrates how that knowledge is used in classification judgments. By claiming that people evaluate category membership on the basis of whether objects were likely to have been generated by a category's causal mechanisms, causal-model theory accounts for not only the differences in feature weights but also for the fact that the particular combination of features displayed by an object affects judgments of category membership. The formalization of causal models offered by causal-model theory advances understanding of the role of knowledge in categorization by enabling quantitative fits to empirical data, producing interpretable parameter estimates, and supporting rigorous tests against other, competing models. As such, causal-model theory helps place the theory-based view of conceptual representation on equal footing with similarity-based models, which attempt to account for human performance without recourse to people's theoretical beliefs.

References

- Ahn, W. (1998). Why are different features central for natural kinds and artifacts?: The role of causal status in determining feature centrality. *Cognition*, *69*, 135–178.
- Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology*, *41*, 361–416.
- Anderson, J. R., & Fincham, J. M. (1996). Categorization and sensitivity to correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 259–277.
- Bloom, P. (1998). Theories of artifact categorization. *Cognition*, *66*, 87–93.
- Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

- Carey, S. (1995). On the origin of causal understanding. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary approach* (pp. 268–302). Oxford: Clarendon Press.
- Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous diagnostic observations. *Journal of Abnormal Psychology, 72*, 193–204.
- Chapman, L. J., & Chapman, J. P. (1969). Illusory correlations as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74*, 272–280.
- Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367–405.
- Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology, 58*, 545–567.
- Chi, M. (1993). Conceptual change within and across ontological categories: Examples from learning and discovery in science. In R. Giere (Eds.), *Cognitive models of science: Minnesota studies in the philosophy of science* (133–190). Minneapolis, Minnesota: University of Minnesota Press.
- Estes, W. K. (1994). *Classification and cognition*. New York: Oxford University Press.
- Gelman, S. A., Coley, J. D., & Gottfried, G. M. (1994). Essentialist beliefs in children: The acquisition of concepts and theories. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind* (pp. 341–367). Cambridge, England: Cambridge University Press.
- Gelman, S. A., & Kalish, C. W. (1993). Categories and causality. In R. Pasnak & M. L. Howe (Eds.), *Emerging theories in cognitive development* (pp. 3–32). New York: Springer-Verlag.
- Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings of the nonobvious. *Cognition, 38*, 213–244.
- Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117*, 227–247.
- Glymour, C. (1998). Learning causes: Psychological explanations of causal explanation. *Minds and Machines, 8*, 39–60.
- Glymour, C., & Cheng, P. W. (1998). Causal mechanism and probability: A normative approach. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 296–313). Oxford, England: Oxford University Press.
- Gopnik, A. (2000). Explanation as orgasm and the drive for causal knowledge: The function, evolution, and phenomenology of the theory-formation system. In F. C. Keil & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 299–324). Cambridge, MA: MIT Press.
- Gopnik, A., & Wellman, H. M. (1994). The “theory theory.” In L. Hirschfeld & S. Gelman (Eds.), *Mapping the mind: Domain specificity in culture and cognition* (pp. 257–293). New York: Cambridge University Press.
- Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior, 18*, 441–461.
- Hayes-Roth, B., & Hayes-Roth, F. (1977). Concept learning and the recognition and classification of examples. *Journal of Verbal Learning and Verbal Behavior, 16*, 321–338.
- Heit, E. (1994). Models of the effects of prior knowledge on category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1264–1282.
- Heit, E., & Bott, L. (2000). Knowledge selection in category learning. In D. L. Medin (Eds.), *The Psychology of Learning and Motivation* (pp. 163–199). New York: Academic Press.
- Jordan, M. I. (Ed.). (1999). *Learning in graphical models*. Cambridge, MA: MIT Press.
- Kaplan, A. S., & Murphy, G. L. (2000). Category learning with minimal prior knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 829–846.
- Keil, F. C. (1979). *Semantic and conceptual development: An ontological perspective*. Cambridge, MA: Harvard University Press.
- Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.
- Keil, F. C. (1995). The growth of causal understandings of natural kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary approach* (pp. 234–262). Oxford: Clarendon Press.
- Lamberts, K. (1995). Categorization under time pressure. *Journal of Experimental Psychology: General, 124*, 161–180.
- Lamberts, K. (1998). The time course of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 695–711.
- Lassaline, M. E. (1996). Structural alignment in induction and similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 754–770.
- Lopez, A., Atran, S., Coley, J. D., Medin, D. L., & Smith, E. E. (1997). The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology, 32*, 251–295.
- Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103–189). New York: Wiley.
- Mackie, J. L. (1974). *The cement of the universe: A study of causation*. New York: Oxford University Press.
- Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior, 23*, 250–269.
- Matan, A., & Carey, S. (2001). Developmental changes within the core of artifact concepts. *Cognition, 78*, 1–26.
- McNamara, T. P., & Sternberg, R. (1983). Mental models of word meanings. *Journal of Verbal Learning and Verbal Behavior, 22*, 449–474.
- Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–196). Cambridge, MA: Cambridge University Press.
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207–238.
- Medin, D. L., & Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology, 20*, 158–190.
- Minsky, M., & Papert, S. (1988). *Perceptrons*. Cambridge, MA: MIT Press.
- Morris, M. W., & Larrick, R. P. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review, 102*, 331–355.
- Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 904–919.
- Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289–316.
- Murphy, G. L., & Wisniewski, E. J. (1989). Feature correlations in conceptual representations. In G. Tiberchien (Ed.), *Advances in cognitive science: Vol. 2. Theory and applications* (pp. 23–45). Chichester, England: Ellis Horwood.
- Neumann, P. G. (1974). An attribute frequency model for the abstraction of prototypes. *Memory & Cognition, 2*, 241–248.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology, 115*, 39–57.
- 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 Kaufman.
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, England: Cambridge University Press.
- Proffitt, J. B., Coley, J. D., & Medin, D. L. (2000). Expertise and category-based induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 811–828.

- Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, 3, 383–407.
- Rehder, B. (1999). A causal model theory of categorization. *Proceedings of the 21st Annual Meeting of the Cognitive Science Society*, Vancouver, British Columbia, Canada, 595–600.
- Rehder, B. (2003). Categorization as causal reasoning. *Cognitive Science*, 27, 709–748.
- Rehder, B., & Burnett, R. C. (2002). Inferring unobserved category features with causal knowledge. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the 24th Annual Meeting of the Cognitive Science Society*, 774–779.
- Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General*, 130, 323–360.
- Rehder, B., & Hastie, R. (in press). Category coherence and category-based property induction. *Cognition*.
- Rehder, B., & Murphy, G. L. (in press). A knowledge-resonance (KRES) model of category learning. *Psychonomic Bulletin & Review*.
- Rehder, B., & Ross, B. H. (2001). Abstract coherent categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 1261–1275.
- Reitman, J. S., & Bower, G. H. (1973). Storage and later recognition of exemplars of concepts. *Cognitive Psychology*, 4, 194–206.
- Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. (pp. 21–59). New York: Cambridge University Press.
- Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, 4, 328–350.
- Rosch, E. H., & Mervis, C. B. (1975). Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology*, 7, 573–605.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 318–362). Cambridge, MA: MIT Press.
- Schank, R. C., Collins, G. C., & Hunter, L. E. (1986). Transcending inductive category formation in learning. *Behavioral and Brain Sciences*, 9, 639–686.
- Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, 75 (Whole No. 517).
- Sloman, S., Love, B. C., & Ahn, W. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, 22, 189–228.
- Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, Mass: Harvard University Press.
- Spalding, T. L., & Murphy, G. L. (1999). What is learned in knowledge-related categories? Evidence from typicality and feature frequency judgments. *Memory & Cognition*, 27, 856–867.
- Sperber, D., Premack, D., & Premack, A. J. (Ed.). (1995). *Causal cognition: A multidisciplinary approach*. New York: Oxford University Press.
- Thomas, R. D. (1998). Learning correlations in categorization tasks using large, ill-defined categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 119–143.
- Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal models and the acquisition of category structure. *Journal of Experimental Psychology: General*, 124, 181–206.
- Wattenmaker, W. D., Dewey, G. I., Murphy, T. D., & Medin, D. L. (1986). Linear separability and concept learning: Context, relational properties, and concept naturalness. *Cognitive Psychology*, 18, 158–194.
- Wisniewski, E. J. (1995). Prior knowledge and functionally relevant features in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 449–468.
- Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science*, 18, 221–282.

Appendix

Configural-Features Prototype Model

Assume that the four-dimensional stimulus space is expanded with three additional dimensions that encode the confirmation or violation of the three causal relationships, $F_1 \rightarrow F_2$, $F_2 \rightarrow F_3$, and $F_3 \rightarrow F_4$. Thus, the category prototype P would be 1111111 (the presence of the fifth, sixth, and seventh feature indicate that the causal relationships are confirmed), and each to-be-classified exemplar would have three additional features indicating which causal relationships are confirmed or violated. According to the configural-features prototype model, the category membership rating assigned to an exemplar E would be equal to its similarity to the category prototype P, scaled by a constant K,

$$\begin{aligned} \text{Rating}(E) &= K \text{Sim}_{\text{prototype}}(E, P) \\ &= K(\Pi_{i=1 \dots 4} S_i)(\Pi_{i=5 \dots 7} Z_i), \end{aligned}$$

where $S_i = 1$ if $E_i = P_i$, otherwise $S_i = s_i$ where $0 \leq s_i \leq 1$, and where $Z_i = 1$ if $E_i = P_i$ otherwise $Z_i = s_v$, where $0 \leq s_v \leq 1$. The free parameters s_1 , s_2 , s_3 , and s_4 are feature weights associated with features F_1 , F_2 , F_3 , and F_4 , respectively, and parameter s_v is the weight associated with violating a causal relationship.

The configural-features prototype model was fit to the category membership ratings of the 36 participants in the chain condition of Experiment

3. For purposes of computing similarity between the three-feature test exemplars of Experiment 3 and the category prototype, the unknown first feature, F_1 , was assumed present, that is, similarity was computed according to an intersection rule (Estes, 1994). Because dimension 1 was not varied (it was always listed as unknown) parameter s_1 was not estimated. Thus, these fits involved estimating five parameters s_2 , s_3 , s_4 , s_v , and K for each participant.

The parameter values averaged over participants were $s_2 = .800$, $s_3 = .846$, $s_4 = .796$, $s_v = .762$, and $K = 94.9$. The fits yielded an average SSE of 854, which is comparable to the 893 achieved by the chain model in Experiment 3 (see Table 5). However, the configural-features prototype model achieves this equivalent fit only at the cost of extra parameters compared with the chain model (five vs. three, respectively). When a measure of fit (RMSD) is used that is sensitive to the number of parameters, the chain model achieves a better average fit compared with the configural-features prototype model (11.7 vs. 14.1, respectively) and a better fit for 27 of the 36 participants.

Received August 2, 2001

Revision received March 17, 2003

Accepted April 14, 2003 ■