ELSEVIER

Cognitive Science 27 (2003) 709–748

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

Categorization as causal reasoning☆

Bob Rehder*

Department of Psychology, New York University, 6 Washington Place, New York, NY 10003, USA

Received 22 October 2002; received in revised form 27 June 2003; accepted 1 July 2003

Abstract

A theory of categorization is presented in which knowledge of causal relationships between category features is represented in terms of asymmetric and probabilistic causal mechanisms. According to *causal-model theory*, objects are classified as category members to the extent they are likely to have been generated or produced by those mechanisms. The empirical results confirmed that participants rated exemplars good category members to the extent their features manifested the expectations that causal knowledge induces, such as correlations between feature pairs that are directly connected by causal relationships. These expectations also included sensitivity to higher-order feature interactions that emerge from the asymmetries inherent in causal relationships. Quantitative fits of causal-model theory were superior to those obtained with extensions to traditional similarity-based models that represent causal knowledge either as higher-order relational features or "prior exemplars" stored in memory.

Keywords: Categorization; Causal reasoning; Causal knowledge; Conceptual representation; Causal models

1. Introduction

One factor that has changed the study of human cognition over the last several decades is the realization that the operation of many cognitive processes depends on the world knowledge that a person possesses. Whereas psychologists traditionally have emulated the simple experiments of physicists' by stripping stimuli of their meaning to discover the "fundamental particles" of

☆ Portions of the empirical findings and the model fitting results were reported previously in a paper to the 2001 NIPS Conference.

* Tel.: +1-212-992-9586; fax: +1-212-995-4349.

E-mail address: bob.rehder@nyu.edu (B. Rehder).

0364-0213/$ – see front matter
doi:10.1016/S0364-0213(03)00068-5

the psychological universe, many now investigate the meaningful explanations, interpretations, understandings, and inferences that world knowledge affords and that characterize much of our everyday mental activity. Sometimes referred to as the *theory-based* view of conceptual representation, an emphasis on the importance of studying the knowledge that people bring to bear to the task at hand now appears in numerous research domains including memory (Bartlett, 1932; Bransford, Barclay, & Franks, 1972), problem solving (Gentner, 1983), language comprehension (Kintsch, 1988), expertise (Chase & Simon, 1973; Chi, Feltovich, & Glaser, 1981), categorization (Murphy & Medin, 1985; Wisniewski & Medin, 1994), judgment (Nisbett & Ross, 1980), learning (Chi, Bassok, Lewis, Reimann, & Glaser, 1989), and cognitive development (Carey, 1985; Gopnik & Meltzoff, 1998; Karmiloff-Smith & Inhelder, 1975; Keil, 1989).

Prior research into the nature of people's intuitive theories suggests that these theories are composed of a number of different types of beliefs. For example, researchers have argued that people's folk theories include ontological beliefs about the types of entities in the world (Chi, 1993; Keil, 1979; Wellman & Gelman, 1992). Carey (1985, 1995) has stressed the distinctive nature of the beliefs involving intentional agents that make people's "naive psychology." Finally, Keil (1995) has proposed that teleological beliefs about the purpose served by various objects and their properties is another fundamental way in which people conceive the world. However, according to many researchers intuitive theories largely consist of beliefs about *causal* relations (Carey, 1995; Gelman, Coley, & Gottfried, 1994; Gopnik & Wellman, 1994; Keil, 1989; Murphy & Medin, 1985). The special status granted causal knowledge is unsurprising in light of its distinct functional advantages. 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).

This article is concerned with the nature of people's intuitive theories about categories of everyday objects, especially within-category knowledge about the causal relationships that link objects' features. Previous research has provided direct evidence that people's knowledge of categories is more extensive than a simple "list of features." For example, Ahn (1998) demonstrated that undergraduates possess extensive causal knowledge about natural categories by having them rate the strength of causal connections between features (e.g., they produced high ratings for "goats give milk because of their genetic code" but low ratings for "goats give milk because they have four legs") (also see Sloman, Love, & Ahn, 1998). Moreover, previous research has also shown that causal knowledge influences the classification of objects. For example, ever since Rosch's emphasis on the importance of the family resemblance structure of natural categories, it has been assumed that the features of a category vary regarding their importance, or *weight*, for categorization (Hampton, 1979; Rosch, 1973; Rosch & Mervis, 1975; Smith & Medin, 1981), and an aim of categorization research has been to understand those factors that make features more or less important to establishing category membership. In fact, research has established a number of robust empirical generalizations regarding how the importance of a feature changes as a function of its position in a network of causal relationships. For example, one generalization is that a feature will become more important to the extent it is more deeply embedded in a causal network of inter-related features (Ahn, 1998; Ahn, Kim, Lassaline, & Dennis, 2000; Rehder, in press; Sloman et al., 1998). Another is that a feature will become more important to the extent it has many causes (Rehder & Hastie, 2001).

Although these findings advance our understanding of the effects of folk theories on classification, there are good reasons for believing that these effects are likely to be more extensive than merely affecting the importance of individual features. After all, the primary role of theories is to inter-relate their constituent features, and hence one might expect that the role that an individual feature plays in establishing category membership will depend on how that feature relates to the other features an object has. In other words, theories might make certain *combinations* of features either sensible and coherent (and other combinations insensible and incoherent) in light of the relations linking them, and the degree of coherence of a set of features might be an important factor determining membership in a category. For example, most adults not only know that typical birds are small, have wings, fly, and build nests in trees, they also know that birds build nests in trees *because* they can fly, and they fly *because* they have a wing size appropriate to their size. In light of this knowledge, a small winged animal that doesn't fly and yet still builds nests in trees might be considered a less plausible bird (how did the nest get into the tree?) than a large winged animal that doesn't fly and builds nests on the ground (e.g., an ostrich), even though the former animal has more features that are typical of birds (three) than the latter (one).

Prior research supports the intuition that an important role of categorizer's prior knowledge is to make particular combinations of features more or less acceptable to category membership. For example, Malt and Smith (1984) found that members of natural categories were seen as more typical when they possessed pairs of correlated properties that were either especially salient or occurred in "functional combinations" (e.g., a large bird requires large wings to fly) [but also see Ahn, March, Luhmann, & Lee, 2002]. Rehder (in press) found that combinations of features that violated a category's causal knowledge (i.e., cause feature absent and effect feature present or vice versa) led to lower category membership ratings (also see Rehder & Hastie, 2001). Wisniewski (1995) reported that certain objects were better examples of the category "captures animals" when they possessed combinations of features that were useful (e.g., "contains peanuts" and "caught an elephant") as compared to when they did not ("contains acorns" and "caught an elephant"). Finally, Rehder and Ross (2001) showed that objects were considered good examples of a category of pollution cleaning devices when they possessed a gathering instrument that was appropriate to the type of pollution being gathered (e.g., "has a metal pole with a sharpened end" and "works to gather discarded paper") but not otherwise ("has a magnet" and "removes mosquitoes").

Despite these demonstrations of the influence of theoretical knowledge on categorization, there have been few attempts to formalize the nature of the causal beliefs that largely make up those theories. One exception is *causal-model theory* (Rehder, 1999, in press; Waldmann, Holyoak, & Fratianne, 1995). According to causal-model theory, people's intuitive theories about categories of objects consist of a model of the category in which both a category's features and the causal mechanisms among those features are explicitly represented. Applied to the bird example above, the claim is that most people's knowledge of birds includes features (wings, small, flies, nests in trees, etc.) interconnected with causal mechanisms (nests in trees because of flight, flight because of wings, and so on). Importantly, causal relations interconnecting features are assumed to be asymmetrical such that one feature (e.g., wings) causes another (e.g., flight) but not vice versa.

According to causal-model theory, people's causal models of categories influence their classification behavior by leading them to expect certain distributions of features in category members. Specifically, a to-be-classified object is considered a category member to the extent that its features were likely to have been *generated* by the category's causal laws, such that combinations of features that are likely to be produced by a category's causal mechanisms are viewed as good category members and those unlikely to be produced by those mechanisms are viewed as poor category members. For example, when two features of a category are believed to be linked by a causal mechanism, those two features will be expected to be correlated with one another—the two features will tend to be both present or both absent in category members. As a result, causal-model theory will assign a low degree of category membership to objects that have many broken feature correlations (i.e., cases where causes are present but effects absent or vice versa). Objects that have many preserved correlations (i.e., causes and effects both present or both absent) will receive a higher degree of category membership because it is just such objects that are likely to be generated by causal laws.

Causal-model theory's predictions that categorizers are sensitive to pairwise correlations between features directly-connected by causal relationships distinguishes it from many existing accounts that consider only the effect knowledge has on the importance of individual features to categorization. However, sensitivity to such pairwise correlations is just one consequence of the claim that categorizers favor those combinations of features that are expected to be generated by a category's causal laws. To illustrate, consider the causal networks presented in Fig. 1. In the *common-cause schema*, one category feature (F_1) is described as causing the three other features (F_2, F_3, and F_4). In the *common-effect schema* one feature (F_4) is described as being caused by each of the three other features (F_1, F_2, and F_3). On the one hand, these schemas are analogs of one another if one ignores the direction of causality. On the other hand, because of the asymmetries inherent in causal relationships the two causal networks shown in Fig. 1 are each expected to generate populations of category members with distinctive distributions of features. Although both schemas imply that causes will be correlated with effects (i.e., the common-cause will be correlated with its effects, and the common-effect will be correlated with its causes), the common-cause network also implies that the three effect attributes will be correlated (because of their common-cause). In contrast, the common-effect schema does *not* imply that the three cause attributes will be correlated. This disanalogy between the pattern of pairwise correlations between features is presented in Fig. 2, and can be illustrated with the following example: Three symptoms caused by a disease (a common-cause network) will be expected to be

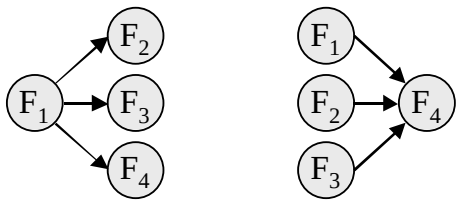

Fig. 1. The common-cause and common-effect causal schemas.

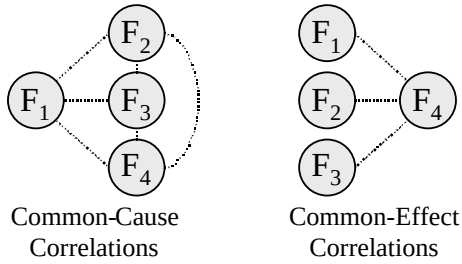

Fig. 2. Feature correlations associated with the common-cause and common-effect causal schemas.

correlated across a population in which the disease is possessed by a subpopulation. In contrast, three independent causes of the disease (a common-effect network) are just that, independent (i.e., uncorrelated). This asymmetry between common-cause and common-effect networks has been the focus of considerable investigation in both the philosophical and psychological literatures (Reichenbach, 1956; Salmon, 1984; Waldmann & Holyoak, 1992; Waldmann et al., 1995).

Nevertheless, there are conditions under which common effect networks will also exhibit statistical structure among category features in addition to pairwise correlations between causally-connected features. Consider the informal description of the higher-order interactions that may hold among features arranged in a common-effect pattern which follows (a more formal description is presented later in this article). First, when the common effect feature F_4 is present in an object, that object will be a better category member if a cause feature (say, F_1) is also present, because the presence of F_1 means that the expected correlation between F_1 and F_4 is preserved. However, the importance of F_1 being present is also likely to depend on whether other causes (F_2 and F_3) are already present—in particular, the presence of F_1 is likely to be less important when there is already one cause present to "explain" the presence of F_4. In other words, a common-effect network predicts higher-order interactions between features such that the influence of the presence a cause feature depends on both the presence of the common effect and its other potential causes, reflecting the diminishing marginal returns associated with explaining an effect that is already adequately explained. The higher-order interactions associated with a common-effect network are related to the normative behavior of discounting for the case of multiple sufficient causation during causal attribution (Morris & Larrick, 1995). Note that these predictions for common-effect networks distinguishes the current proposal from the account presented by Waldmann et al. (1995) who suggest that higher-order structure is absent in common-effect networks. However, their analysis is based on continuous variables and structural equation modeling, and I will demonstrate below how analysis based on discrete variables (e.g., binary category features), indeed predicts the higher-order correlations among common-effect features just described.

The goal of this article is to test causal-model theory's predictions regarding the influence of causal knowledge on judgments of category membership. To this end, undergraduate participants were instructed on categories whose four binary features exhibited either a common-cause or a common-effect schema. Table 1 presents an example of the features and common cause causal relationships for one of the novel categories used in this study, Lake Victoria Shrimp.

Table 1
Features and common cause causal relationships for the Lake Victoria Shrimp experimental category

	Features
F_1	High amounts of ACh neurotransmitter
F_2	Long-lasting flight response
F_3	Accelerated sleep cycle
F_4	High body weight
	Causal relationships
$F_1 \rightarrow F_2$	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
$F_1 \rightarrow F_3$	A high quantity of the ACh neurotransmitter causes an accelerated sleep cycle. The neurotransmitter speeds up all neural activity, including the internal "clock" which puts the shrimp to sleep on a regular cycle.
$F_1 \rightarrow F_4$	A high quantity of the ACh neurotransmitter causes a high body weight. The neurotransmitter stimulates greater feeding behavior, which results in more food ingestion and more body weight.

After learning about either a common-cause or common-effect category, categorization ratings were gathered from experimental participants. To determine participants' sensitivity to features, pairs of features, and higher-order interactions between features, multiple-regressions were performed on these ratings. If categorizers are insensitive to causal asymmetries, the common-cause and common-effect schemas should result in analogous classification behavior because these schemas are analogs of one another if one ignores the direction of causality. On the other hand, if categorizers are sensitive to the pattern of data generated by causal relations their categorization performance should exhibit a disanalogy between common-cause and common-effect schemas, a disanalogy that arises from the inherent asymmetries in causal relationships—that causes generate their effects but not vice versa.

The following experiment will demonstrates that, as predicted by causal-model theory, categorizers are sensitive to whether to-be-classified exemplars preserve or break the pairwise and higher-order inter-feature correlations that are expected to inhere in a population of category members. Moreover, beside confirming these predictions of causal-model theory in qualitative terms, in the following section I present a formal definition of causal-model theory and show that the theory also provides a good quantitative account of categorization performance.

Although causal-model theory will be shown to provide a sufficient account of categorization performance, it is also important to consider whether such performance can be accounted for by extensions to more traditional models such as similarity-based prototype and exemplar models. Although similarity-based models have been shown to characterize categorization behavior in numerous experimental studies over the last several decades, these models were not primarily intended to account for categorization performance in the presence of categorizers' intuitive theories. However, natural extensions to these models intended to account for the effects of such theories have been proposed, and considerable theoretical parsimony would be achieved if these extensions could account for categorization performance in light of those theories. Thus, a second purpose of the current article is to consider whether such proposals are sufficient to account for the categorization performance in light of causal knowledge that links features of categories.

2. Method

2.1. Materials

Six novel categories were constructed: two biological kinds (Kehoe Ants, Lake Victoria Shrimp), two non-living natural kinds (Myastars, Meteoric Sodium Carbonate), and two artifacts (Romanian Rogos, Neptune Personal Computers). Each category had four binary features which were described as distinctive relative to a superordinate category. Each feature was described as probabilistic, that is, not all category members possessed it (e.g., "Some Lake Victoria Shrimp have a high quantity of the ACh neurotransmitter whereas others have normal amounts." "Some Lake Victoria Shrimp have a fast flight response whereas others have a normal flight response," etc.). Throughout this article the presence of a feature is denoted with "1," and its absence with "0." Each causal relationship was described as one feature causing another (e.g., "A high quantity of ACh neurotransmitter causes a long-lasting flight response."), accompanied with one or two sentences describing the causal mechanism (e.g., "The duration of the electrical signal to the muscles is longer because of the excess amount of neurotransmitter."). A complete description of the categories' cover story, features, and causal relationships is available at http://cogsci.psy.utexas.edu/supplements/.

2.2. 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 first presented with the cover story and the category's features and their base rates. Participants in the common-cause condition were then instructed on the common-cause causal relationships ($F_1 \rightarrow F_2$, $F_1 \rightarrow F_3$, and $F_1 \rightarrow F_4$), participants in the common-effect condition were instructed on the common-effect relationships ($F_1 \rightarrow F_4$, $F_2 \rightarrow F_4$, and $F_3 \rightarrow F_4$), and control participants were told of no causal links between features. Common-cause and common-effect participants also observed a diagram like those in Fig. 1 depicting the structure of the causal links. When ready, all participants took a multiple-choice test that tested them on the knowledge they had just studied. During the test participants could request help in which case the computer re-presented the information about the category. However, participants were required to retake the test until they committed 0 errors and made 0 requests for help. For common-cause and common-effect participants the test consisted of 21 questions. For control participants the test consisted of 7 questions.

All participants then performed three tasks counterbalanced for order: a classification task, a property induction task, and a similarity rating task. The results from the property induction and similarity task are unrelated to the theoretical issues raised in this report and are omitted. During the classification task, participants rated the category membership of 48 exemplars, consisting of all possible 16 objects that can be formed from four binary features and the eight single-feature exemplars, each presented twice. For example, those participants assigned to learn the Lake Victoria Shrimp category were asked to classify a shrimp that possessed "High amounts of the ACh neurotransmitter," "A normal flight response," "An accelerated sleep cycle," and "Normal body weight." The feature values of each to-be-rated exemplar

were listed in order (F_1 through F_4) on the computer screen. The list of feature values for single-feature exemplars contained "???" for the three unknown features. The order of the 48 exemplars was randomized for each participant.

Participants entered their rating with a response bar that appeared underneath the exemplar. The left and right arrow keys were used to move a bar along a horizontal scale to a position which reflected confidence in the exemplar's category membership. The left end of the scale was labeled "Definitely not an X" and the right end was labeled "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 the range 0–100. Experimental sessions lasted for approximately 45 minutes.

2.3. Participants

One hundred and eight University of Illinois undergraduates received course credit for participating in this experiment. They were randomly assigned in equal numbers to the three conditions, and to one of the six experimental categories.

3. Results

Category membership ratings for the 16 test exemplars averaged over participants in the common-cause, common-effect, and control conditions are presented in Table 2, and in Fig. 3 for selected exemplars. The presence of inter-feature causal relationships in the common-cause

Table 2
Observed categorization ratings, and the ratings predicted by causal-model theory in the common-cause and common-effect conditions

	Common cause		Common effect		Control
Exemplar	Observed	Predicted	Observed	Predicted	Observed
0000	60.0 (5.3)	61.7	70.0 (4.8)	69.3	70.7 (3.3)
0001	44.9 (3.8)	45.7	26.3 (4.0)	27.8	67.0 (3.1)
0010	46.1 (3.9)	45.7	43.4 (3.8)	47.7	65.6 (3.5)
0100	42.8 (3.7)	45.7	47.3 (4.0)	47.7	66.0 (3.4)
1000	44.5 (5.1)	44.1	48.0 (3.7)	47.7	67.0 (3.6)
0011	41.0 (4.0)	40.1	56.3 (3.6)	56.5	67.1 (3.1)
0101	40.8 (4.0)	40.1	56.5 (3.8)	56.5	66.5 (3.5)
0110	42.7 (4.1)	40.1	38.3 (3.8)	39.2	65.6 (3.3)
1001	55.1 (4.1)	52.7	57.7 (3.4)	56.5	68.0 (2.9)
1010	52.6 (4.3)	52.7	43.0 (4.0)	39.2	67.6 (3.3)
1100	54.3 (3.7)	52.7	41.9 (3.6)	39.2	69.9 (2.8)
0111	39.4 (4.6)	38.1	71.0 (2.9)	74.4	67.6 (3.1)
1011	64.2 (3.8)	65.6	75.7 (2.1)	74.4	67.2 (2.9)
1101	65.3 (4.0)	65.6	74.7 (2.0)	74.4	70.2 (2.9)
1110	62.0 (4.1)	65.6	33.8 (5.7)	35.8	72.2 (2.7)
1111	90.8 (2.4)	89.6	91.0 (2.2)	90.0	75.6 (2.7)

Standard errors for observed ratings are shown in parentheses.

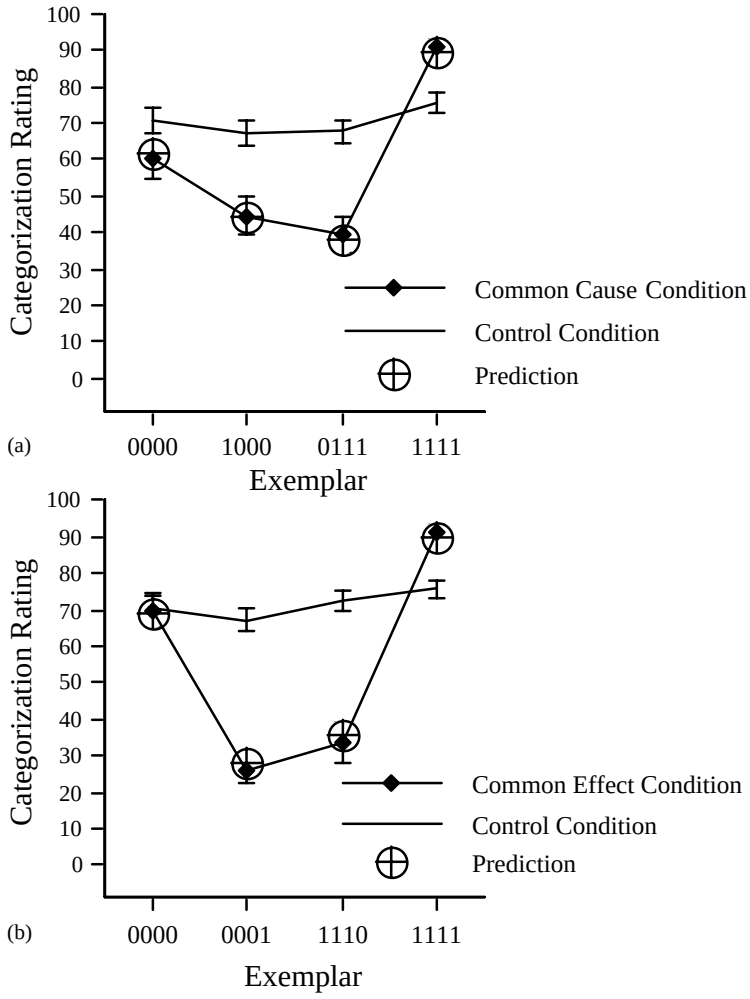

Fig. 3. Categorization ratings for selected exemplars. (a) In the common-cause and control conditions. (b) In the common-effect and control conditions. Ratings predicted by causal-model theory are shown in each panel.

and common-effect conditions had a large effect on the category membership ratings relative to the control condition. For instance, exemplars 1000 and 0111 were given significantly lower ratings in the common-cause condition (44.5 and 39.4, respectively) than in the control condition (67.0 and 67.6). In addition, exemplars 0001 and 1110 were given significantly lower ratings in the common-effect condition (26.3 and 33.8, respectively) than in the control condition (67.0 and 72.2). Presumably, these lower ratings in the common-cause and common-effect conditions arose because those exemplars broke expected pairwise correlations between features. For example, in the common-cause condition exemplar 1000 breaks all three expected correlations because features F_2, F_3, and F_4 are absent even though their cause F_1 is present, and exemplar 0111 breaks all three correlations because features F_2, F_3, and F_4 are present even though F_1 is absent. Similarly, in the common-effect condition exemplar 1110 breaks

expected correlations because features F_1, F_2, and F_3 are present even though their effect F_4 is absent, and exemplar 0001 breaks all three correlations because features F_1, F_2, and F_3 are absent even though F_4 is present.

Fig. 3 also indicates that exemplar 1111 received a significantly higher rating in the common-cause (90.8) and common-effect (91.0) conditions as compared to the control condition (75.6). The higher rating received by 1111 in the causal schema conditions obtained presumably because in both conditions exemplar 1111 preserves all three expected pairwise correlations (cause and effect features are both present for all three causal links). Overall, the pattern of results shown in Fig. 3 indicates the presence of interactions between features because categorization ratings were a non-monotonic function of the number of features: test exemplars with 0 and 4 features (i.e., 0000 and 1111) received higher ratings than exemplars with one and three features (i.e., 1000 and 0111 in the common-cause condition, 0001 and 1110 in the common-effect condition).

An effect of causal schema condition on category membership ratings was confirmed by statistical analysis. A two-way ANOVA with condition (3 levels: common-cause, common-effect, control) and exemplar (16 levels) with repeated measures on the second factor was conducted. In this analysis there was a significant interaction between experimental condition and exemplar ($F(28, 1470) = 12.77$, MSE $= 294.5$, $p < .0001$) reflecting the fact that the pattern of ratings given to exemplars differed in the three conditions. More specifically, the pattern of categorization ratings differed between the common-cause and control conditions (significant interaction between the common-cause/control contrast and the exemplar factor, $F(14, 1470) = 8.16$, $p < .0001$), and between the common-effect and control conditions (interaction between the common-effect/control contrast and exemplar, $F(14, 1470) = 17.11$, $p < .0001$).

It was suggested earlier that causal schemas might affect the importance of interactions among features, especially those that preserve or break expected correlations among features. To test these hypotheses directly, category membership ratings were analyzed by performing a multiple regression for each participant. First, four predictor variables (f_1, f_2, f_3, f_4) were coded as -1 if the feature was absent, and $+1$ if it was present. The regression weight associated with each f_i represents the influence the feature had on category membership ratings. Positive weights will result if the presence of a feature increases the categorization rating and if the absence of the feature decreases it. Second, an additional six predictor variables were constructed by computing the multiplicative interaction between each pair of features: f_{12}, f_{13}, f_{14}, f_{24}, f_{34}, and f_{23}. The resulting interaction terms are coded as -1 if one of the features is present and the other absent, and $+1$ if both are present or both absent. For those feature pairs on which a causal relationship is defined, the two-way interaction terms represent whether an expected pairwise correlation is preserved ($+1$, cause and effect both present or both absent) or broken (-1, cause present and effect absent, or cause absent and effect present). The regression weights associated with these two-way interactions index the influence that the preservation or breaking of the correlation had on categorization ratings, with a positive weight indicating that confirmation leads to a higher categorization rating and violation to a lower one. Finally, the four three-way interactions (f_{123}, f_{124}, f_{134}, and f_{234}) were also included as predictors in the per-participant regressions to test for the presence of higher-order interactions among features. The regression weights averaged over participants in the common-cause, common-effect, and control conditions are presented in Fig. 4. I first discuss the impact of

causal knowledge first on the weights of individual features and then on interactions between features.

3.1. Feature weights

Fig. 4 demonstrates that causal knowledge affected the weight given to individual features. In the common-cause condition, the common-cause (f_1) carried greater weight (8.2) than the

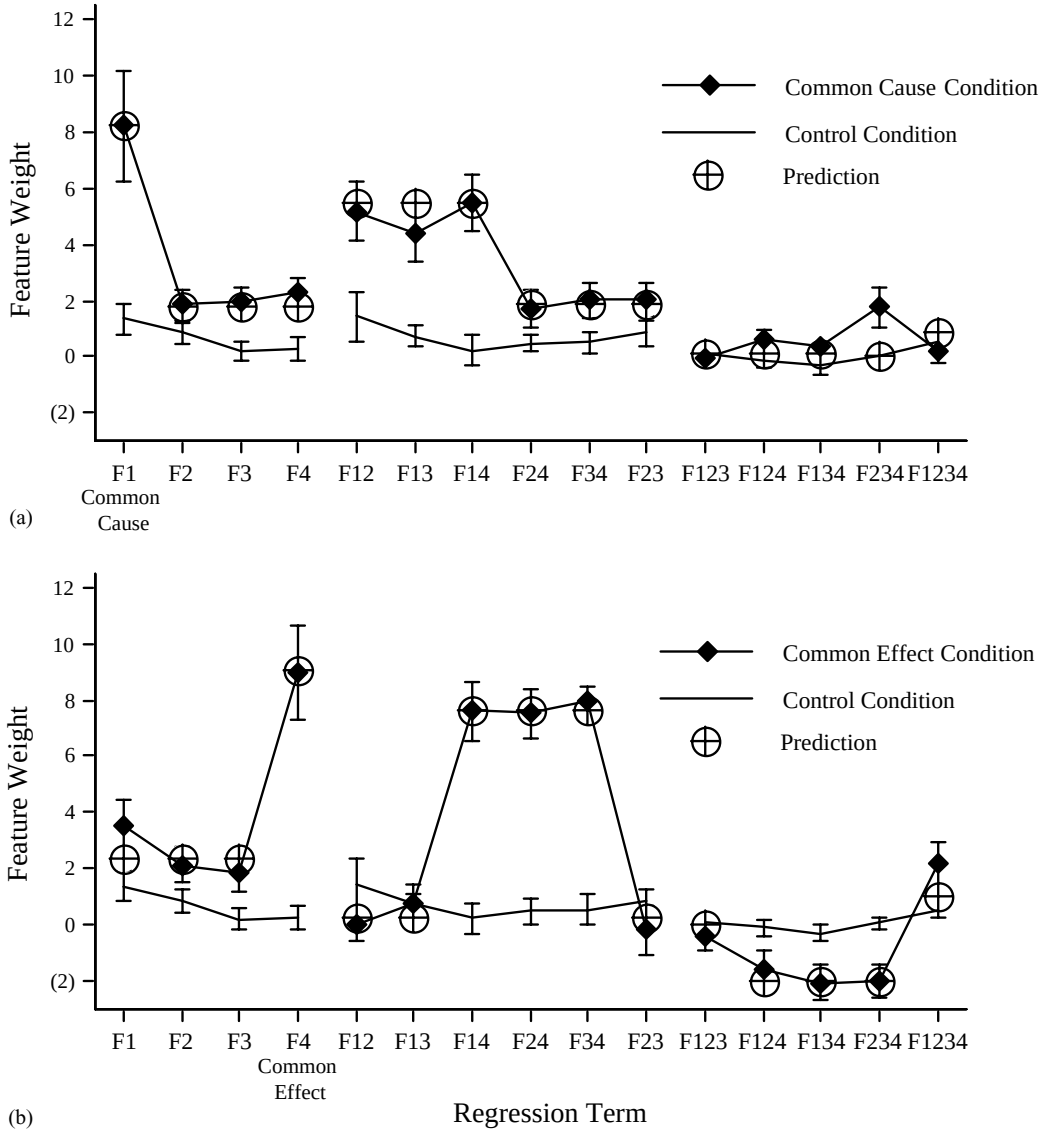

Fig. 4. Regression weights for (a) the common-cause vs. control conditions, and (b) the common-effect vs. control conditions. Weights predicted by causal-model theory are also shown in each panel.

three effects (1.8, 1.9, and 2.3 for f_2, f_3, and f_4). And, in the common-effect condition the common-effect (f_4) had greater weight (9.0) than the three cause features (3.5, 2.1, and 1.9 for f_1, f_2, and f_3, respectively). These differences in feature weights were confirmed by statistical analysis. A 3 (causal schema) by 4 (feature: 1, 2, 3 or 4) ANOVA on the regression weights was carried out, with repeated measures on the last factor. The pattern of feature weights differed as a function of causal schema (as indicated by a significant interaction between schema and feature, $F(6, 315) = 9.33$, MSE $= 26.3$ $p < .0001$). More specifically, the pattern of weights differed between the common-cause and control conditions ($F(3, 315) = 4.83$, $p < .005$), and between the common-effect and control conditions ($F(3, 315) = 8.35$, $p < .0001$). The greater importance of any feature that participates in many causal relationships such as common causes and common effects replicates past research (Rehder & Hastie, 2001).

3.2. Feature interactions

Fig. 4 also demonstrates that causal knowledge affected the weight given to pairwise interaction terms. In particular, the pairwise interaction terms corresponding to those feature pairs assigned causal relationships had positive weights in both the common-cause condition (5.1, 4.3, and 5.5 for f_{12}, f_{13}, and f_{14}), and the common-effect condition (7.6, 7.5, and 8.0 for f_{14}, f_{24}, and f_{34}). That is, category membership ratings were higher when a cause and effect feature and were either both present or both absent, and lower when one of those features was absent and the other present. This result indicates that participants considered an exemplar a better category member when it preserved the category's expected correlations and a worse member when it broke those correlations. As expected, in the control condition the six two-way interaction terms were all approximately zero.

A 3 (causal schema) by 6 (two-way interaction term) ANOVA on the regression weights confirmed that the pattern of weights given to the two-way interaction terms depended on the causal schema (significant interaction between causal schema and interaction term, $F(10, 525) = 26.72$, MSE $= 10.3$, $p < .0001$). The pattern of two-way interactions in the common-cause and common-effect conditions each differed from the control condition ($F(5, 525) = 4.95$, $p < .0001$; $F(5, 525) = 34.59$, $p < .0001$). The three two-way interaction weights corresponding to the three causal relationships differed from the control condition (all p's $< .005$), in both the common-cause and common-effect conditions.

The results presented thus far are consistent with the hypothesis that the common-cause and common-effect schemas have analogous effects on category membership ratings. First, in both causal schemas the feature common to many causal relationships (i.e., the common cause and common effect) became most heavily weighted in category membership decisions. Second, in both schemas participants attended to whether cause and effect features were in agreement with one another. However, earlier in this article I argued that causal-model theory predicts an important disanalogy between these two causal schemas, a prediction that was shown to follow from the asymmetry of causal relationships. In fact, the regression weights shown in Fig. 4 confirms that each condition exhibited a unique pattern of feature interactions, a pattern characteristic of the associated causal network. First, it was shown earlier (Fig. 2) that the common-cause schema implies that the three effect features will be correlated, because of their common-cause. Consistent with this prediction, in the common-cause condition the three

two-way interaction terms between the effect features (f_{24}, f_{34}, f_{23}) are significantly greater than those interactions in the control condition (1.7, 2.0, and 2.0 vs. 0.4, 0.5, and 0.8, $F(1, 70) = 4.08$, MSE $= 7.44$, $p < .05$). In contrast, the common-effect schema does not imply that the three cause features will be correlated, and in fact in the common effect condition the interactions between the cause attributes (f_{12}, f_{13}, f_{23}) did not differ from those in the control condition (see Fig. 4).

Second, the common-effect schema produces a pattern of classification behavior that results in not just interactions between pairs of features, but also higher-order feature interactions. In that condition interaction terms f_{124}, f_{134}, f_{234}, and f_{1234} were -1.6, -2.0, -2.0, and 2.2

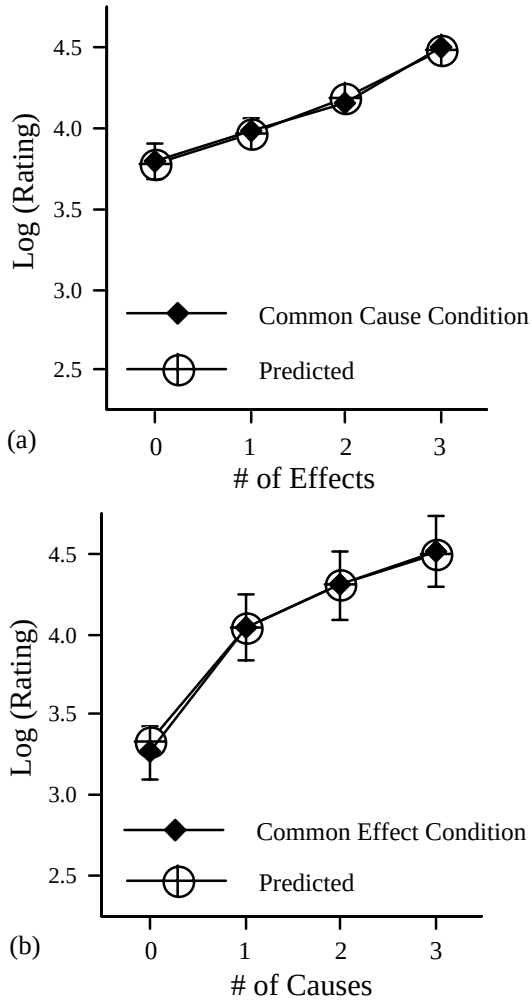

Fig. 5. The logarithm of observed categorization ratings: (a) in the common-cause condition when the common cause is present as a function of the number of effect features; (b) in the common-effect condition when the common effect is present as a function of the number of causes. Ratings predicted by causal-model theory are shown in each panel.

as compared to −0.1, −0.3, 0.1, and 0.5 in the control condition. A 3 (causal schema) by 5 (higher-order interaction term) ANOVA on the regression weights confirmed that the pattern of weights given to the higher-order interaction terms depended on the causal schema (significant interaction between causal schema and interaction term, $F(8, 420) = 3.15$, MSE = 8.1, $p < .05$). In particular, the pattern of higher-order interaction terms differed between the common-effect and control conditions ($F(4, 420) = 4.96$, $p < .005$). In the common-cause condition causal-model theory predicts no higher-order interactions, and, indeed, the higher-order interactions terms in that condition did not differ from those in the control condition ($F < 1$).

I argued earlier that higher-order interactions are expected from a common-effect schema because it predicts a non-linear increase in category membership ratings as the number of cause features present to "explain" the common effect feature increases—in particular, categorization ratings are expected to experience a greater increase when the number of cause features increases from zero to one as compared to when additional cause features are introduced. Fig. 5b presents the logarithm of categorization ratings in the common-effect condition for those exemplars in which the common effect is present as a function of the number of cause features. As predicted, categorization ratings increase more with the introduction of the first cause feature as compared to subsequent cause features. (The reason for log coordinates in Fig. 5 is explained below.) Apparently, participants considered the presence of at least one cause explaining the presence the common-effect to be sufficient grounds to grant an exemplar a relatively high degree of category membership in a common-effect category.

In contrast, Fig. 5a shows a linear increase in (the logarithm of) category membership ratings for those exemplars in which the common cause is present as a function of the number of effect features. That is, in the presence of the common cause, each additional effect produced a constant increment in the goodness of category membership.

4. Discussion

According to causal-model theory, judgments of category membership are based on how likely an object was to have been generated by a category's causal laws. One sign of the operation of a causal law between two features is the presence of a pairwise correlation between those features. As predicted, the introduction of causal relationships in both a common-cause pattern and a common-effect pattern resulted in participants increasing their category membership ratings when an object's features preserved those expected pairwise correlations, and decreasing their ratings when the features broke them.

It was also argued earlier that the asymmetrical nature of causal relations is such that they can be expected to produce additional structure in category members. As predicted, common-cause but not common-effect participants were sensitive to pairwise correlations between effect features in the common-cause condition. In addition, common-effect but not common-cause participants were sensitive to higher-order correlations between features that were consistent with a strategy that required a common-effect to have one cause to explain its presence. Taken together, the pattern of feature interactions found in this experiment reflects the application of qualitatively different classification strategies in the common-cause

and common-effect conditions. These results demonstrate that undergraduates are able to perceive causal links as more than simple associations, and take into account the asymmetries inherent in causal relationships (Waldmann & Holyoak, 1992; Waldmann et al., 1995).

The current experiment also assessed the importance of features considered individually. As discussed earlier, Ahn and her colleagues (Ahn, 1998; Ahn et al., 2000; Ahn & Lassaline, 1995; Sloman et al., 1998) have advanced specific proposals regarding the effects of causal knowledge on feature weights, namely that features are more important to categorization to the extent that they are "more causal," that is, more deeply embedded in a causal network. In contrast to these proposals, the common-effect results indicate that an effect feature that has many potential causes becomes heavily weighted. This finding replicates previous research with common-effect networks (Rehder & Hastie, 2001).

In the section that follows, I present a formal definition of causal-model theory, and then fit the theory to the present categorization data. As will be shown, causal-model theory not only provides a qualitative account of the current results, but also an excellent quantitative account.

5. Causal-model theory

The central claim of causal-model theory is that people's knowledge of categories includes not just features but also a representation of the causal mechanisms that link those features. Fig. 6 demonstrates a simple causal model in which one feature, C, is depicted as the cause of a second feature, E. Causal models such as those presented in Figs. 1 and 6 are specialized instances of Bayesian networks (Glymour, 1998; Jordan, 1999; Pearl, 1988). In this article the variables are binary category features whose values represent whether the feature is present or absent (values also denoted by "1" and "0," respectively).

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 non-descendent variable when the state of its parent variables are known). However, by itself the network conveys no information regarding the details of the causal relationships that link variables in a network. In contrast, causal-model theory makes specific assumptions regarding the functional form of the causal relationships between binary variables, namely, it assumes that people view features as being linked by probabilistic causal mechanisms. It is assumed that when a cause feature is present (e.g., C in Fig. 6) it enables the operation of a causal mechanism that will, with some probability, bring about the presence of the effect feature (e.g., E in Fig. 6). When the cause feature C is absent it is assumed that it has no causal influence on the effect E.

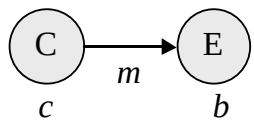

Fig. 6. A simple causal model with two binary features and one causal relationship.

The probabilistic nature of the causal mechanism linking C and E is represented by parameter *m*. Parameter *m* is the probability that the probabilistic mechanism that links C and E will successfully operate (i.e., will bring about the presence of E) when C is present. Parameter *m* corresponds to a probabilistic version of the familiar notion of the *sufficiency* of causal relations. Causal sufficiency obtains whenever a cause is always accompanied by its effect, and can be represented by setting *m* equal to 1. Causal models also allow for the possibility that there might be other unspecified causes of effects. Parameter *b* represents the probability that E will be present even when it is not brought about by C. Parameter *b* corresponds to a probabilistic version of the familiar notion of the *necessity* of causal relations. Causal necessity obtains whenever an effect is always accompanied by its cause, and can be represented by setting the effect's *b* parameter equal to 0. Parameter *b* can be interpreted as the probability that E is brought about by some unspecified background cause other than C. Finally, parameter *c* represents the probability that feature C will be present.

An important characteristic of this formalization of causal knowledge is that it reflects our intuitive understanding of the asymmetry of causal relationships. First, it is assumed that when the causal mechanism presumed to link C and E operates (with probability *m*) it generates E but not C. Second, it is assumed that the mechanism potentially operates only when C is present but never when C is absent. As a result, causal-model theory is sensitive to which values of the binary variables C and E in Fig. 1 are called "present" and which are called "absent," and also which variable is called the "cause" and which the "effect." In contrast, a symmetrical relation like correlation is insensitive to which value of a binary variable is labeled present and which is labeled absent and also to whether the variable plays the role of cause or effect.

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 likelihood that the model will generate any combination of the features can be expressed as a function of the model's parameters. For example, Table 3 presents the likelihoods that the causal model of Fig. 6 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 (i.e., $P(\sim C \sim E)$, 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 (i.e., $P(\sim CE)$ or $P(01)$) is $(1 - c)$ times the probability that E is brought about by any background cause, *b*. The probability that C is present but E absent (i.e., $P(C \sim E)$ or

Table 3
Likelihood equations for the causal model of Fig. 2

Exemplar (E)	$L(E; c, m, b)$
00	$[1 - c][1 - b]$
01	$[1 - c][b]$
10	$[c][(1 - m)(1 - b)]$
11	$[c][m + b - mb]$

$P(10)$) is c times the probability that E is not brought about the causal mechanism *and* not brought about by the background cause $(1 - m)(1 - b)$. Finally, the probability that C and E are both present (i.e., $P(CE)$ or $P(11)$) is c times the probability that E is brought about by the causal mechanism *or* brought about by the background cause $(m + b - mb)$, or equivalently $[1 - (1 - m)(1 - b)]$. 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 0–1.

In this article I will assume that the causal links of the common-cause and common-effect schemas shown in Fig. 1 are each assumed to be constituted by a probabilistic causal mechanism like that assumed for the simple model of Fig. 6. Specifically, in the common-cause model it is assumed that three probabilistic causal mechanisms link F_1 with F_2, F_3, and F_4. These three causal mechanisms are assumed to operate independently and each with probability m. Likewise, in the common-effect model it is assumed that three probabilistic causal mechanisms that operate independently each with probability m link F_1, F_2, F_3 with F_4. In each model it is also assumed that each effect feature has potential background causes (b for F_2, F_3, and F_4 in the common-cause model, and b for F_4 in the common-effect model) that operate independently. Finally, each cause feature has a parameter representing the probability that it will be present (c for F_1 in the common-cause model, and c for F_1, F_2, and F_3 in the common-effect model).[1]

Likelihood equations for the common-cause and common-effect models can be derived by applying the same Boolean algebra operations that were applied to the simple model of Fig. 6. For example, the probability of exemplar 1101 (i.e., F_1, F_2 and F_4 present, F_3 absent) being generated by a common-cause model is the probability that F_1 is present [c], times the probability that F_2 was brought about by F_1 or its background cause $[m + b - mb]$, times the probability that F_3 was brought about by neither F_1 nor its background cause $[(1 - m)(1 - b)]$, times the probability that F_4 was brought about by F_1 or its background cause $[m + b - mb]$. Likewise, the probability of exemplar 1011 being generated by a common-effect model is the probability that F_1 is present [c], times the probability that F_2 is absent $[1 - c]$, times the probability that F_3 is present [c] times the probability that F_4 was brought about by F_1, F_3, or its background cause $[1 - (1 - m)(1 - m)(1 - b)]$. The 16 possible combination of values for F_1, F_2, F_3, and F_4 for the common-cause and common-effect models are presented in Table 4 as a function of the parameters c, m, and b.

As described earlier, a fundamental claim of causal-model theory is that although causal mechanisms are reflected in feature correlations, they are not equivalent to those correlations. In particular, the common-cause and common-effect causal schemas were selected for testing in this research in order to demonstrate that causal relationships exhibit an asymmetry that is not characteristic of the symmetric relation of correlation. This lack of equivalency between pairwise causal mechanisms and pairwise feature correlations can be illustrated by examining the likelihoods of exemplars expected to be generated by the common-cause and common-effect models each instantiated with parameter values $c = 0.6$, $m = 0.5$, and $b = 0.3$. (The qualitative predictions about to be described do not depend on these specific values, but rather hold for all values of these parameters >0 and <1.) The resulting exemplar likelihoods are presented in Table 4. From these likelihoods the correlations between features that would obtain in populations of category exemplars that perfectly instantiated those distributions were calculated. As expected, pairs of features linked by causal relationships are correlated with one another: For the common-cause schema there are correlations between the common-cause

Table 4
Likelihood equations for the common-cause (CC) and common-effect (CE) models

Exemplar (E)	L_{CC}(E; c, m, b)	L_{CC}(E; 0.6, 0.5, 0.3)	L_{CE}(E; c, m, b)	L_{CE}(E; 0.6, 0.5, 0.3)
0000	$[1-c][1-b][1-b][1-b]$	0.137	$[1-c][1-c][1-c][1-b]$	0.045
0001	$[1-c][1-b][1-b][b]$	0.059	$[1-c][1-c][1-c][b)$	0.019
0010	$[1-c][1-b][b][1-b]$	0.059	$[1-c][1-c][c][(1-m)(1-b)]$	0.034
0100	$[1-c][b][1-b][1-b]$	0.059	$[1-c][c][1-c][(1-m)(1-b)]$	0.034
1000	$[c][(1-m)(1-b)][(1-m)(1-b)][(1-m)(1-b)]$	0.026	$[c][1-c][1-c][(1-m)(1-b)]$	0.034
0011	$[1-c][1-b][b][b]$	0.025	$[1-c][1-c][c][1-(1-m)(1-b)]$	0.062
0101	$[1-c][b][1-b][b]$	0.025	$[1-c][c][1-c][1-(1-m)(1-b)]$	0.062
0110	$[1-c][b][b][1-b]$	0.025	$[1-c][c][c][(1-m)(1-m)(1-b)]$	0.025
1001	$[c][(1-m)(1-b)][(1-m)(1-b)][m+b-mb]$	0.048	$[c][1-c][1-c][1-(1-m)(1-b)]$	0.062
1010	$[c][(1-m)(1-b)][m+b-mb][(1-m)(1-b)]$	0.048	$[c][1-c][c][(1-m)(1-m)(1-b)]$	0.025
1100	$[c][m+b-mb][(1-m)(1-b)][(1-m)(1-b)]$	0.048	$[c][c][1-c][(1-m)(1-m)(1-b)]$	0.025
0111	$[1-c][b][b][b]$	0.011	$[1-c][c][c][1-(1-m)(1-m)(1-b)]$	0.119
1011	$[c][(1-m)(1-b)][m+b-mb][m+b-mb]$	0.089	$[c][1-c][c][1-(1-m)(1-m)(1-b)]$	0.119
1101	$[c][m+b-mb][(1-m)(1-b)][m+b-mb]$	0.089	$[c][c][1-c][(1-(1-m)(1-m)(1-b))]$	0.119
1110	$[c][m+b-mb][m+b-mb][(1-m)(1-b)]$	0.089	$[c][c][c][(1-m)(1-m)(1-m)(1-b)]$	0.019
1111	$[c][m+b-mb][m+b-mb][m+b-mb]$	0.165	$[c][c][c][1-(1-m)(1-m)(1-m)(1-b)]$	0.197

attribute and its effects ($r = .343$), and for the common-effect schema there are correlations between the common effect and its causes ($r = .197$). However, although the common-cause schema produces correlations between the three effects ($r = .118$), the common-effect schema does not produces correlations between the three causes ($r = 0$). These correlations conform to the pattern of pairwise feature correlations presented earlier in Fig. 2.

Additional analysis of this example reveals that there is also higher-order structure among features in the common-effect case. Earlier I argued on intuitive grounds that a common-effect schema predicts that category membership ratings should experience a greater increase when the number of cause features increases from zero to one as compared to when additional cause features are introduced, but that no such effect should obtain for the common-cause network. In the common-cause case, the likelihood of an exemplar that has the common cause feature as a function of the number of effect features also present is 0.026, 0.048, 0.089 and 0.165 for 0, 1, 2, and 3 effect features, respectively. In the common-effect example, the likelihood of an exemplar that has the common effect feature as a function of the number of cause features also present is 0.019, 0.062, 0.119, and 0.197 for 0, 1, 2, and 3 cause features, respectively. Note that the common-effect likelihoods experience a greater increase as the number of cause features increases from zero to one (from 0.019 to 0.062) than do the common-cause likelihoods as the number of effect features increases from zero to one (from 0.026 to 0.048). Moreover, it can be shown that when logarithms of the likelihoods are taken, the common-cause log likelihoods experience a constant increase (of 0.269) as the number of effect features increases. In contrast, the common-effect log likelihoods experience a large increase (of 0.519) when the first cause feature is added to explain the presence of the common-effect, but that this increase decreases as additional cause features are added (0.280 for the second cause, 0.220 for the third). That is, a *discounting effect* is predicted in which there are diminishing marginal returns expected for explaining an effect that is already adequately explained.

The discounting effect distinguishes the current development of causal-model theory from that presented by Waldmann et al. (1995). Waldmann et al. derive their predictions regarding the statistical structure of category members assuming continuous variables and applying structural equation modeling. For example, the relationship between a continuous common effect and three continuous causes is defined to be,

$$E = w_1C_1 + w_2C_2 + w_3C_3 + U$$

where w_i is the weight representing the strength of cause C_i, and U represents a random error component. However, this formalization of causal relations does not predict a discounting effect. Thus, the discounting effect represents a novel prediction of the current formalization of causal relations based on binary variables.

5.1. Theoretical modeling of empirical results

To assess whether the common-cause and common-effect causal models provide a quantitative in addition to a qualitative account of the patterns of classification found in the current experiment, those models were fitted to the category membership ratings of each participant in the common-cause and common-effect conditions. Note that 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 choice axiom to predict choice probabilities. For example, the probability that an exemplar E will be classified into category A versus B would be given by

$$P(\mathrm{A}|\mathrm{E}) = \frac{L_{\mathrm{A}}(\mathrm{E})}{[L_{\mathrm{A}}(\mathrm{E}) + L_{\mathrm{B}}(\mathrm{E})]}$$

where L_{A} and L_{B} are the likelihoods that E was generated by A and B's causal models, respectively (see Rehder, 1999 for an application). 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. Hence, each participant's ratings were predicted from the equation,

$$\text{Rating}\,(\mathrm{E}) = KL(\mathrm{E};\, c, m, b)$$

where $L(\mathrm{E};\, c, m, b)$ is the likelihood of exemplar E as a function of c, m, and b. The likelihood equations for the common-cause and common-effect models shown in Table 4 were used for common-cause and common-effect participants, respectively. K is a scaling constant that brings the likelihoods into the range 0–100. For each of the 36 participants in the common-cause condition, the set of common-cause model parameters K, c, m, and b that minimized the squared deviation between predictions and observations was computed. Likewise, the set of common-effect model parameters K, c, m, and b that minimized the squared deviation between predicted and observed ratings was computed for each of the 36 common-effect participants. The best fitting values for parameters K, c, m, and b averaged over participants are presented in Table 5 for the two conditions.

The significantly positive estimate for m in both conditions indicates how causal-model theory accounts for participants' sensitivity to correlations among features by assuming the presence of probabilistic causal laws. For example, earlier it was shown how common-cause participants gave lower category membership ratings to exemplars 1000 and 0111, and a higher rating to exemplar 1111. Fig. 3a presents the fit of the common-cause model to these exemplars, and indicates that it provides an excellent fit to each. According to causal-model theory, exemplars 1000 and 0111 are poor category members because they are unlikely to be generated by common-cause causal laws (because they break many expected correlations), and 1111 is a good category member because it is likely to be generated by those laws (because it preserves many expected correlations). Likewise, Fig. 3b indicates how the common-effect model provides an excellent fit in the common-effect condition to the low ratings assigned to exemplars that broke many expected correlations (0001 and 1110) and to the high rating assigned to the exemplar that preserved many correlations (1111).

To understand the predictions of causal-model theory in terms of feature weights and interactions between features, each participant's predicted ratings were subjected to the same multiple regression that was performed on the observed ratings. The resulting regression weights averaged over participants are presented in Fig. 4 superimposed on the weights from the observed data. Fig. 4 confirms that causal-model theory reproduces participants' sensitivity to agreement between pairs of features directly connected by causal relationships in both the common-cause condition (f_{12}, f_{13}, and f_{14}) and the common-effect condition (f_{14}, f_{24}, and f_{34}). Moreover, Fig. 4 indicates that causal-model theory is also able to account for those interaction terms distinctive to each causal model. That is, it accounts for the interactions between the effect

Table 5
Model fitting results

Parameters	Common-cause condition	Common-effect condition
Causal-model theory		
c	0.578 (0.020)	0.522 (0.005)
m	0.214 (0.042)	0.325 (0.042)
b	0.437 (0.019)	0.280 (0.033)
K	846 (42.9)	876 (30.2)
Avg. RMSD	8.7	12.1
Configural-Features Prototype Model		
s_c	0.779 (0.043)	0.935 (0.016)
s_e	0.946 (0.012)	0.749 (0.038)
s_v	0.792 (0.042)	0.782 (0.026)
K	90.6 (2.5)	99.9 (2.4)
Avg. RMSD	8.8	14.8
Exemplar-Fragments Model		
s_c	0.389 (0.063)	0.456 (0.047)
s_e	0.499 (0.066)	0.258 (0.058)
g	0.334 (0.056)	0.357 (0.051)
K	61.1 (4.3)	63.9 (4.3)
Avg. RMSD	10.1	15.7

Standard errors for parameter estimates are shown in parentheses. RMSD = root mean square error.

features in the common-cause condition (f_{23}, f_{24}, and f_{34}). It also accounts for the higher-order feature interactions in the common-effect condition (f_{124}, f_{134}, f_{234}, and f_{1234}). The presence of these distinctive patterns of feature interactions in the two causal schemas indicates that causal-model theory, like the undergraduate participants, is sensitive to the asymmetry inherent in causal relationships. Fig. 5 presents the predicted category membership ratings in the presence of the common cause (Fig. 5a) and the common effect (Fig. 5b) as a function of the number of effect and cause features, respectively. As the figure indicates, causal-model theory reproduces the observed linear increase in ratings (in log coordinates) as one adds effect features to a common cause, and the observed non-linear increase in ratings as one adds cause features to a common effect (i.e., the discounting effect). Finally, Fig. 4 indicates that causal-model theory is able to account for the increased weight associated with the common cause in the common-cause condition, and with the common-effect in the common-effect condition.

6. Similarity-based accounts of correlated features

The previous section demonstrate that causal-model theory predicts and experimental participants exhibit a sensitivity to correlated features in classification judgments as a result of the presence of causal knowledge about a category. Nevertheless, it is important to also consider

whether such sensitivity can be accounted for by extensions to more traditional categorization models such as similarity-based prototype and exemplar models. As argued earlier, considerable theoretical parsimony would be achieved if these extensions could account for classification performance in light of causal knowledge. In this section the fits of causal-model theory to the classification data will be compared to fits achieved by extensions of similarity-based prototype and exemplar models that have been proposed to account for the effects of prior knowledge on categorization.

6.1. Prior knowledge and prototype models

As discussed earlier, some researchers have proposed that the effects of prior knowledge on categorization can be characterized in terms of changes to subjective feature weights (Ahn, 1998; Ahn et al., 2000; Keil, 1989; Medin & Shoben, 1988). However, because a well-known property of prototype models is that categorization is not influenced by combinations of features above and beyond the features individually (Kemler Nelson, 1984; Medin, Altom, Edelson, & Freko, 1982; Medin & Schwanenflugel, 1981), a prototype model is in principle unable to account for the sensitivity to feature interactions reported in the present experiment. Nevertheless, there remains the possibility that a form of prototype model might be able to account for the results by noting that prototype models leave open the question of what counts as a feature. For example, one might postulate the existence of second-order features (Gluck & Bower, 1988; Hayes-Roth & Hayes-Roth, 1977; Minsky & Papert, 1988; Murphy, 1993; Neumann, 1974; Reitman & Bower, 1973; Rumelhart, Hinton, & Williams, 1986; Wattenmaker, Dewey, Murphy, & Medin, 1986) that encode whether expected correlations are preserved or broken, and that participate in similarity computations alongside primitive features. To illustrate, in the common-cause condition exemplar 1110 might include three second-order properties which indicate that the first and second correlations are preserved (F_1 and F_2 and F_3 all present) but that the third relationship is broken (F_1 present but F_4 absent). A category membership judgment would then be made on the basis of the number of matching features between this 7-feature exemplar (1110110, with the second-order features 110 coded on the fifth, sixth, and seventh dimensions) and a 7-feature category prototype 1111111 that represents that the prototypical category member has all four features present and preserves all three expected correlations. The result of this comparison would be a match on three of the first four dimensions and two of the last three. In contrast, the exemplar 0111 would be encoded as 0111000 because it breaks all three expected correlations, and although it matches the category prototype 1111111 on three of the first four dimensions, it mismatches on the last three. Thus, even though both exemplars possess the same number of primitive features (three), 0111 would receive a lower category membership rating than 1110 because of its greater number of missing second-order features (i.e., broken correlations).

This proposal, which I shall refer to as the *Configural-Features Prototype Model*, is formalized in Appendix A. The Configural-Features Prototype Model was fit to the category membership ratings of each of the participants in the common-cause and common-effect conditions. These fits involved estimating four parameters s_c, s_e, s_v, and K for each participant, where s_c is the weight associated with a cause feature (F_1 in the common-cause condition, F_1, F_2, and F_3 in the common-effect condition), s_e is the weight associated with an effect feature

(F_4 in the common-effect condition, F_2, F_3, and F_4 in the common-cause condition), s_v is the weight associated with a broken correlation, and K is a scaling constant. The best fitting values for these parameters averaged over participants in the common-cause and common-effect conditions are presented in Table 5. Note that lower estimates for a weight parameter means that the feature is more influential; an estimate of 1 means that the presence or absence of the feature has no influence on categorization decision.

As Table 5 indicates, in both common-cause and common-effect conditions parameter s_v differed significantly from 1, reflecting the sensitivity to pairwise correlations between features directly connected by causal relationships. For example, in the common-cause condition the fits to exemplars 0000, 1000, 0111, and 1111 were 60.6, 43.6, 38.5, and 90.6, respectively. The sensitivity to correlated features is reflected in the lower estimates for exemplars with many broken correlations (1000 and 0111) as compared to higher estimates for exemplars with many preserved correlations (0000 and 1111). Likewise, in the common-effect condition the fits to exemplars 0000, 0001, 1110, and 1111 were 61.3, 39.5, 38.3, and 99.9, respectively, and sensitivity to correlated features is indicated by the lower estimates for those exemplars with broken correlations (0001 and 1110) versus the higher estimates for those exemplars with preserved correlations (0000 and 1111).

Table 5 also indicates an effect of causal schemas on the weight of individual features. In the common-cause condition, the common cause feature F_1 was more influential ($s_c = 0.779$) than the corresponding effect features ($s_e = 0.946$). Similarly, in the common-effect condition, the common effect feature F_4 was more influential ($s_e = 0.749$) than the corresponding cause features ($s_c = 0.935$). These differences in model weights mirror the pattern of regression-based feature weights presented earlier (Fig. 4).

These results indicate that the Configural-Features Prototype Model is able to account for the effects of causal knowledge on the importance of individual features and of pairwise correlations between features directly connected by causal relationships. However, as I have stressed repeatedly, because causality is an asymmetric relationship it produces statistical structure among category features in addition to pairwise feature correlations, such as higher-order interactions involving the common-effect and its causes in a common-effect network. Fig. 7a and b present the fits of the Configural-Features Prototype Model to the data presented earlier in Fig. 5. Fig. 7b indicates that the Configural-Features Prototype Model is unable to account for the discounting effect. That is, because the Configural-Features Prototype Model limits its account of causal knowledge to an effect of whether expected pairwise correlations are confirmed or violated, it is unable to capture higher-order interactions among features predicted by a common-effect network.

This failure of the Configural-Features Prototype Model is reflected in measures of the degree of fit of the model to the data relative to causal-model theory. The root mean square deviation (RMSD) associated with each models' fits was computed for each common effect participant.[2] Causal-model theory yielded a better average RMSD than the Configural-Features Prototype Model (12.1 vs. 15.7), and a better fit (according to RMSD) for 25 of 36 participants.

Fig. 7a indicates that, in contrast to the common-effect condition, the Configural-Features Prototype Model does a good job of accounting for the common-cause ratings for those exemplars in which the common cause is present as a function of the number of effect features. One reason for this result was that the higher-order features defined by the Configural-Features

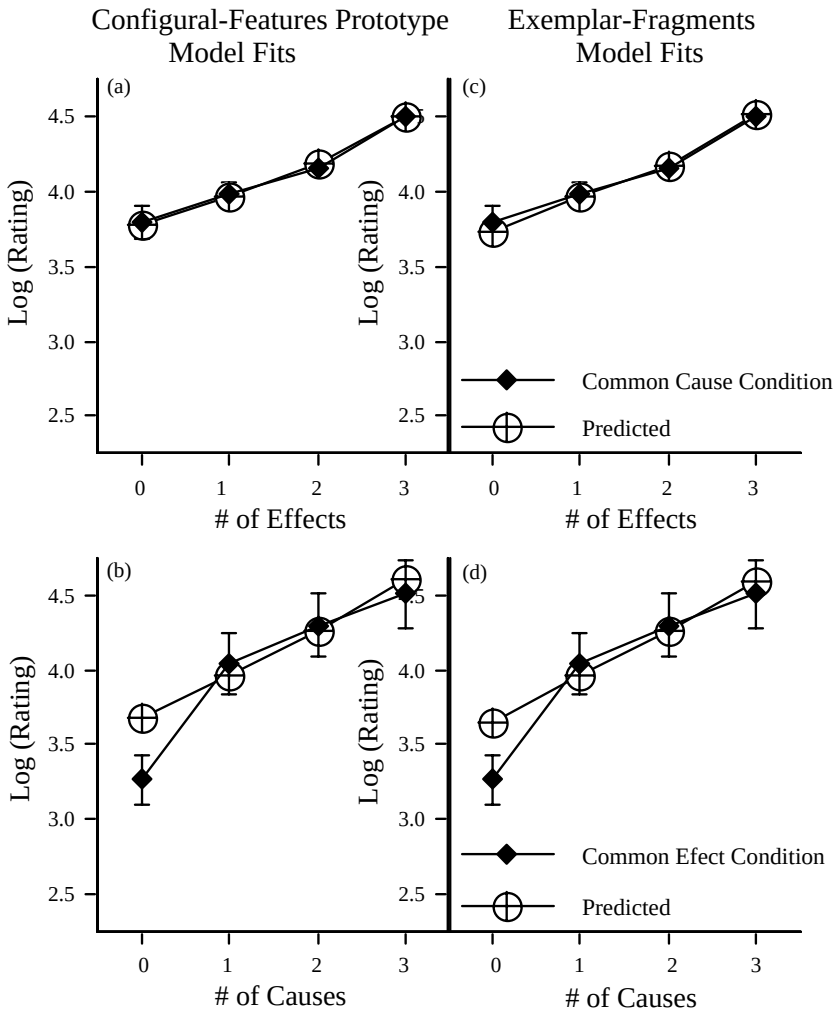

Fig. 7. Model fits for the Configural-Features Prototype and Exemplar-Fragment Models.

Prototype Model accounts for correlations between features directly connected by causal relationships. Another is that the Configural-Features Prototype Model assumes a multiplicative similarity rule, such that predicted ratings are a non-linear function of the features (both basic and higher-order) that an object displays. The result of this non-linear function is that the model also exhibits a sensitivity to correlations between effect features. For example, regression analyses run on the predictions of the Extended-Features Prototype Model in the common-cause condition yielded an average weight on the two-way interactions between effect features (f_{23}, f_{24}, and f_{34}) of 1.8, which is in good agreement with the empirical weights on these terms of 1.9 (see Fig. 4a). However, because the Extended-Features Prototype Model attributes this sensitivity to correlations among effect features in the common-cause condition merely to participant's logarithmic use of the response scale, it also predicts a sensitivity to pairwise

correlations between cause features in the common-effect condition. Indeed, regression analyses run on the predictions of the Extended-Features Prototype Model in the common-effect condition yielded an average weight on the two-way interactions between cause features (f_{12}, f_{13}, and f_{23}) of 1.6. However, no such sensitivity is found in that condition's empirical results (average weights on those terms of 0.2, see Fig. 4b). Thus, the pattern of two-way interactions in the common-cause condition cannot be attributed merely to participant's use of a logarithmic response scale.

6.2. Prior knowledge and exemplar models

Another way to try to account for the categorization results in terms of a similarity-based model is to extend an exemplar model such as the context model (Medin & Schaffer, 1978; Nosofsky, 1986). The context model predicts sensitivity to correlated features, although ordinarily this sensitivity emerges as a by-product of classifying by similarity to previously-observed category exemplars stored in memory. However, in this study participants observed no exemplars of the category, and as a result any sensitivity to correlated features must be attributed directly to the causal knowledge that participants learn rather than stored exemplars. Nevertheless, Heit (1994, 1998) has proposed that exemplar models can be extended to accommodate the effects of prior knowledge by assuming that such knowledge takes the form "prior exemplars" stored in memory. On this account, categorization involves computing the similarity between a to-be-classified exemplar and stored category exemplars, including those exemplars that represent prior knowledge.

To account for the effects of causal knowledge on categorization in this section I consider a prototype model elaborated with a number of partial exemplars representing causal knowledge stored in memory. The partial exemplars encode the inter-feature correlations expected from causal relations. For example, in the common-cause condition I will assume that participants encode the common-cause causal relationships as three pairs of partial exemplars: 11*xx* and 00*xx* ("*x*" representing an unknown value) representing the $F_1 \rightarrow F_2$ causal link, 1*x*1*x* and 0*x*0*x* representing the $F_1 \rightarrow F_3$ link, and 1*xx*1 and 0*xx*0 representing the $F_1 \rightarrow F_4$ link. Likewise, in the common-effect condition I assume that the common-effect causal links are encoded as 1*xx*1 and 0*xx*0 (for link $F_1 \rightarrow F_4$), *x*1*x*1 and *x*0*x*0 (for link $F_2 \rightarrow F_4$), and *xx*11 and *xx*00 (for link $F_3 \rightarrow F_4$). These partial exemplars enable an exemplar model to exhibit sensitivity to pairwise feature correlations, because an exemplar that breaks many correlations will be dissimilar to those stored exemplars. For example, in the common-cause condition although exemplar 0111 matches three of the four features of the category prototype 1111, it perfectly matches the features of none of six partial exemplars 11*xx*, 00*xx*, 1*x*1*x*, 0*x*0*x*, 1*xx*1, and 0*xx*0. In contrast, exemplar 1110 matches not only three of the four features of the category prototype, but it also perfectly matches the features of two of the partial exemplars (11*xx* and 1*x*1*x*). Thus, all else begin equal, 1110 will receive a higher category membership rating than 0111.

This proposal, which I shall refer to as the *Exemplar-Fragments Model*, is formalized in Appendix B. The Exemplar-Fragments Model was fit to the category membership ratings of each of the participants in the common-cause and common-effect conditions. These fits involved estimating four parameters s_c, s_e, g, and K for each participant, where s_c and s_e represent

the weight associated with cause and effect features, respectively (as in the Configural-Features Prototype Model), K is a scaling constant, and g represents the relative weight given to the similarity of a new to-be-classified exemplar to the exemplar fragments versus the prototype.

The best fitting values for parameter s_c, s_e, g, and K averaged over participants in the common-cause and common-effect conditions are presented in Table 5. Consistent with the regression-based feature weights presented earlier (Fig. 4) and the fits of causal-model theory and the Configural-Features Prototype Model, the fits of the Exemplar-Fragments Model reveal an effect of causal schemas on the weight of individual features. In the common-cause condition, the common cause feature F_1 had more influence ($s_c = 0.389$) than the corresponding effect features ($s_e = 0.499$), and in the common-effect condition, the common effect feature F_4 had more influence ($s_e = 0.258$) than the corresponding cause features ($s_c = 0.456$). In addition, in both common-cause and common-effect conditions parameter g differed significantly from 0, reflecting the sensitivity to pairwise correlations between features directly connected by causal relationships. This sensitivity is demonstrated by the fits to exemplars that either preserve or break many correlations: In the common-cause condition the fits to exemplars that broke many expected correlations (42.4 for 1000 and 39.1 for 0111) were lower than the fits to exemplars that preserved many correlations (52.5 for 0000 and 91.5 for 1111). Similarly, in common-effect condition the fits to exemplars that broke many expected correlations (38.3 for 0001 and 36.6 for 1110) were lower than the fits to exemplars that preserved many correlations (55.5 for 0000 and 98.2 for 1111). These fits capture at a qualitative level the effects of preserved and broken correlations seen in the original empirical data (Fig. 3).

However, like the Configural-Features Prototype Model, the Exemplar-Fragments Model represents causal knowledge as a symmetric relation. Fig. 7c and d presents the fits of the Configural-Features Prototype Model to the data presented earlier in Fig. 5. As Fig. 7d indicates, the Exemplar-Fragments Model is unable to account for the common-effect categorization ratings for those exemplars in which the common effect is present as a function of the number of cause features. Like the Configural-Features Prototype Model, it is unable to account for discounting effect. This failure of the Exemplar-Fragments Model in the common-effect condition is reflected in a measure of the degree of fit of the model (Avg. RMSD = 15.7) relative to causal-model theory (Avg. RMSD = 12.1), and the fact that is produced a worse fit than causal-model theory for 24 of the 36 common effect participants.

7. General discussion

The current results support the claim that people have a representation of the causal mechanisms that link category features, and that they categorize by evaluating whether an object's features were likely to have been generated by those mechanisms. That is, people have models of the world that lead them to expect a certain distribution of features in category members, and consider exemplars good category members to the extent they manifest those expectations.

A key assumption of causal-model theory is that the presence of causal knowledge changes one's expectations regarding not only individual features, but also the entire combination of features that a category member is likely to display. This result was predicted on the grounds

that a primary role of theoretical knowledge is to determine how well the features of an object relate to, or cohere with, one another. As predicted, participants considered good category members those exemplars whose combination of features were likely to be generated by the category's causal laws, such as those which preserved correlations between features directly connected by causal relationships. In addition, participants were also sensitive to correlations between effect features in a common-cause schema, and to higher-order interactions among causes and their common effect in a common-effect schema. These additional results indicate that participants were sensitive to even quite subtle aspects of the statistical structure of exemplars that are generated by a category's causal model. In the following section I discuss implications of causal-model theory's generative approach to computing evidence for category membership, and the asymmetrical and probabilistic representation of causality on which it is based.

The model fitting results revealed that not only did causal-model theory provide a good qualitative account of these findings, it provided a good quantiative account as well. These quantitative fits also enabled comparison with similarity-based prototype and exemplar models augmented with extensions to represent the causal knowledge that was presented to participants. The result was that causal-model theory achieved better fits to the categorization ratings, and that both similarity-based models were unable to account for important qualitative trends in the data. In the section following the next I discuss the apparent difficulties associated with accounting for categorization decisions on the basis of overlapping features when theoretical knowledge is present.

Finally, it also important to note that although this article emphasizes the influence that causal knowledge has on determining the importance of combinations of features, causal knowledge also influenced the importance of features individually. In the final section, I discuss the finding that causal knowledge increases the importance of common-cause and common-effect features relative to other features.

7.1. Asymmetric and probabilistic causal mechanisms

Since Hume the traditional analysis of causal relations has been that they are indistinguishable from the symmetric relation of correlation. This tradition has largely continued in modern cognitive psychology. For example, the study of causal reasoning and attribution has often assumed that the perception of causality is no more than the perception of correlation under special conditions, namely, temporal precedence (causes must not follow their effects in time), spatial contiguity, and certain assumptions regarding the default background conditions against which causal induction takes place (Cheng & Novick, 1990, 1991, 1992; Kelly, 1973; Schustack & Sternberg, 1981).

However, more recent research supports the view that people's knowledge of causal relationships differs in important ways from the knowledge of the corresponding correlations. For example, Waldmann and Holyoak (1992) found that learners exhibited cue competition effects when predicting the presence of an effect from multiple possible causes but not when predicting a cause from multiple possible effects (also see Waldmann, 2000; Waldmann et al., 1995). Ahn, Kalish, Medin, and Gelman (1995, Experiment 4) found that a statement of causal mechanism between two types of events had greater influence on subsequent attributions than

did a statement of covariation between those events, even though the expressed magnitude of covariation was derived from the causal mechanism itself. Cheng's (1997) power PC theory of causal induction rests on the assumption that although people generally induce the strength of a causal relationship from the observed co-occurrences between causes and effects, dissociations between judgments of causal strength and correlations arise under certain boundary conditions. Finally, theorists have noted that causal versus correlational knowledge is critical for successfully intervening in the world (manipulating causes produces their effects but not vice versa) and for reasoning about counterfactual situations (Pearl, 2000; Sperber et al., 1995). The current result of disanalogous classification with common-cause and common-effect networks extends the importance of distinguishing between causality and correlation to the domain of categorization, because those two schemas are indistinguishable from one another if one ignores the direction of the causal arrow.

The representation of causal laws offered by causal-model theory incorporates the principle that causality-is-not-correlation by proposing that cause and effect features are linked by causal mechanisms. This representation of causal knowledge is inherently asymmetric because it is the cause that produces the effect rather than the effect producing the cause (and rather than the absence of the cause producing the absence of the effect). Moreover, this representation of causality was key to the success of causal-model theory's generative approach to calculating an exemplar's degree of category membership, because it predicts that populations of category members will exhibit statistical structure in addition to pairwise correlations between causes and effects. In particular, causal-model theory predicts—and people apparently expect—that a common-cause network produces correlations between effects in addition to those between the effects and the cause. And, both causal-model theory and people expect that a common-effect network produces higher-order interactions among the causes and the effect. The current finding that participants considered good exemplars to be exactly those that matched these expectations provides strong support for causal-model theory's generative approach to classification on the basis of asymmetric causal laws.

Another important assumption of causal-model theory is that the causal mechanisms that link category features are viewed as operating probabilistically rather than deterministically. This assumption received direct support from the parameter values derived from the quantitative model fitting. Parameter m, the probability that the causal mechanism will produce its effect, is a probabilistic version of *causal sufficiency*, where deterministic causality (i.e., the cause is always sufficient to bring about its effect) appears as a limiting case when $m = 1$. In fact, across all participants the average value of m was only 0.270, and causal model fits for only 4 of 72 participants yielded values of m greater 0.8. Another assumption of causal-model theory is that people will allow for the possibility that effect features might have causes other than those explicitly stated in a category's causal schema. Parameter b corresponds to a probabilistic version of *causal necessity* because when $b = 0$ an effect is always accompanied by (at least one of) its causes. In fact, the average value of parameter b was 0.359, and the model fits of only 12 of the 72 participants yielded a value of b less than 0.20. The fact that participants assigned an exemplar some significant chance of category membership even if a cause was present and its effect absent, or if a cause was absent and its effect present, supports the claim that people typically treat causal laws probabilistically rather than as a relation of deterministic necessity and sufficiency.

Other theorists have proposed asymmetrical and probabilistic representation of causal knowledge. For example, Rehder (in press) has shown that the representational assumptions of causal-model theory are the same as those implicit in Cheng's (1997) power PC theory of causal induction (e.g., the current *m* parameter corresponds to Cheng's notion of *causal power*; also see Glymour & Cheng, 1998). Working with continuous variables, Waldmann et al. (1995) has derived some of the same predictions presented here (correlations between effects in a common-cause network, no correlations among causes in a common-effect network), and found that the difficulty of learning was greatly influenced by whether the pattern of correlations among features matched the causal model they were led to expect. However, the structural equation modeling approach adopted by Waldmann et al. does not predict the discounting effect found with binary variables in the present common-effect condition and which has been so instrumental in distinguishing causal-model theory from alternative models.

The predictions of causal-model theory have been borne out for a variety of causal networks. First, although the current experiment used what Waldmann et al. (1995) have referred to as a *varying* common-cause schema (because not all category members possess the common-cause), I have also tested a *fixed* common-cause schema in which all category exemplars possess the common-cause. The importance of the common cause being fixed rather than varying lies in the statistical structure of generated exemplars: correlations among the effects that arise when the common cause is varying will be absent when it is fixed.[3] An additional experiment testing this prediction (reported at http://cogsci.psy.utexas.edu/supplements/) replicated the first except that F_1 in the common-cause condition, and F_4 in the common-effect condition, were described as occurring in 100% of all category members. As predicted, in the fixed common-cause condition participants treated effect features as if they were independent, that is, their category membership ratings exhibited no sensitivity to correlations among the effects. Also as predicted, in the fixed common-effect condition participants continued to exhibit a discounting effect such that the presence of the first cause feature was taken as stronger evidence in favor of the common-effect (and hence category membership) than additional causes.

Second, I have instructed participants on categories where the binary features are arranged in a causal chain (i.e., F_1 causes F_2, which causes F_3, which causes F_4) (Rehder, in press). As predicted, participants were sensitive to the pattern of correlations expected to be generated by a causal chain. First, they were strongly sensitive to those feature pairs that should be strongly correlated (those directly connected by causal links: F_1 and F_2, F_3 and F_4, and F_3 and F_4). Second, they exhibited weaker sensitivity to those pairs that should be weakly correlated in a causal chain (those indirectly connected by causal links: F_1 and F_3, F_2 and F_4, and F_1 and F_4).

In the current experiment participants were expected to treat the causal mechanisms between pairs of features as independent of one another, an interpretation encouraged by the presentation of each causal link in a separate paragraph with its own information about mechanism. This situation is presented schematically in Fig. 8a, where diamonds represent the location of causal mechanisms. However, alternative materials could have led participants to assume that each schema included mechanisms involving more than two features. For example, in Fig. 8b, the common cause feature enables a single mechanism that operates probabilistically but that produces all three effects when it operates (e.g., a single genetic mutation causes multiple birth defects). Analogously, the common effect is produced by a single mechanism that requires all

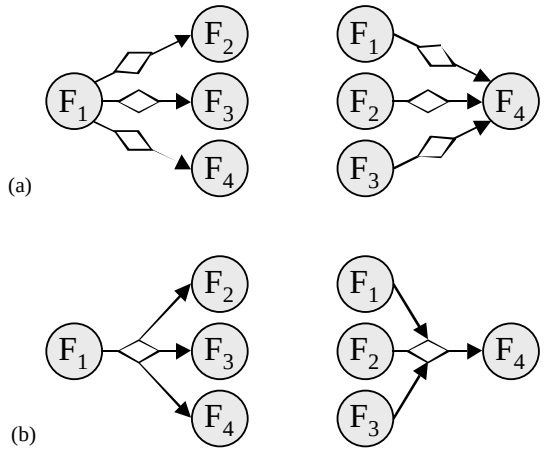

Fig. 8. Alternative interpretations of common cause and common effect models.

three causes to be present to operate (e.g., oxygen, fuel, and spark together produce fire[4]). Once the pattern of causal mechanisms linking three or more category features is established, it is straightforward to apply the equations of causal-model theory, such as to the schemas shown in Fig. 8b.

7.2. Categorization by causal reasoning versus similarity

An important goal of the current research was to determine if the effects of causal knowledge on categorization could be accommodated within the framework of similarity-based models. Two possibilities were considered. First, in the Configural-Features Prototype Model information about whether a to-be-classified exemplar preserved or broke expected correlations between features connected by causal links was encoded as the presence or absence of features on extra dimensions. Second, in the Exemplar-Fragments Model those expected correlations were encoded in the form of "prior" exemplars stored in memory to which to-be-classified exemplars are compared. However, both of these proposals for representing causal knowledge assume that the effect of that knowledge can be fully accounted for in terms of symmetric correlations between features directly-connected by causal relations. As a result, these models were unable to account for the effects of causal knowledge attributed to the asymmetries inherent in causal relations, such as the higher-order interactions involving cause features in the common-effect condition.

Of course, besides the Configural-Features and Exemplar-Fragments Models there are many other proposals that could be considered for how causal knowledge might be represented as higher-order features, or as prior exemplars stored in memory, and it is possible that one of these alternative proposals might successfully account for the current results. For example, although regression analyses were used in this article only as an analytical device to characterize the empirical results, one might consider the possibility that each significantly non-zero regression weight corresponds to one "feature," and that categorizers generated their category membership ratings by matching on those features. However, besides being clearly *post hoc*,

the problem with such a proposal is that it provides no principled explanation for why the set of "features" used in the common-cause conditions (the four basic features plus two-way interactions between features) should differ from those used in the common-effect condition (which in addition found three- and four-way interactions between features).

Alternatively, one might imagine that categorizers mentally generate a representative sample of category members from their causal models, and then judge category membership by computing the similarity of a new exemplar to that sample. On this account, differences between the common-cause and common-effect models arise because their respective mentally-generated samples manifest different patterns of inter-feature correlations. But although such an account might serve as a process model for how causal-model theory's likelihood equations are implemented (or approximated) the bulk of the explanatory burden is being carried by the cognitive structures that generate the mental samples, namely, the causal models themselves.

Rather than searching for representations of causal knowledge that would allow an interpretation of the current categorization results in terms of similarity, I suggest that a more fruitful approach is to view classification as sometimes involving mental processes other than computing a weighted sum of the number of matching features. In this article I have suggested that the current results should be characterized as a case of *categorization as causal reasoning*. On this account, higher-order interactions in, say, a common-effect network are attributed to reasoning processes that tell us that the first explanation of a common-effect is more important than subsequent explanations. Other studies demonstrating the influence of prior knowledge on classification may be usefully understood as instances of causal reasoning. For example, Wisniewski (1995), found that participants rated artifacts with novel combinations of features (e.g., "contains peanuts" and "caught an elephant") as good examples of an animal-capturing device, presumably because of causal reasoning processes that informed them of the usefulness of a trap that contains bait that is attractive to the prey as compared to one that does not ("contains acorns" and "caught an elephant"). Similarly, Rehder and Ross (2001) found that objects with completely novel features (e.g., "has a metal pole with a sharpened end" and "works to gather discarded paper") were rated as good examples of a novel type of pollution cleaning device presumably because of causal reasoning processes that informed them of the usefulness of a device that possesses an instrument to gather the type of pollution as compared to one that does not ("has a magnet" and "removes mosquitoes") (for other likely cases of causal reasoning during categorization, see Heit, 2001; Murphy & Allopenna, 1994; Pazzani, 1991; Wattenmaker et al., 1986).

One outstanding question concerns how causal models come to be associated with categories in the first place. In some cases people's causal models may come from external sources such as formal education. This may be especially true for scientific concepts for which direct observations are rare or impossible (e.g., subatomic particles, galaxies, viruses, etc.). In other cases, causal models may be triggered by the first few category members one encounters. For example, Rehder and Ross (2001) suggested that initial examples of categories are often not just observed but rather *comprehended*, that is, they spontaneously elicit and combine with knowledge a person already possesses to construct a new category representation. On this account, the influence of empirical observations on classification is mediated by the knowledge structures those observations elicit, and the causal reasoning processes that those structures subsequently enable (also see Heit & Bott, 2000; Rehder & Murphy, in press).

Once a causal model for a category has been established, another important question concerns the nature of the interaction between that model and the subsequent observation of category members. One possibility is that the causal classification processes proposed here and traditional similarity-based processes operate side-by-side, each making an independent contribution to the ultimate categorization decision (Wisniewski & Medin, 1994). Another is that the information carried by the observations is integrated into the causal model by the tuning of its parameters. For example, the *b* parameter may be tuned on the basis of observations in which an effect occurs in the absence of a cause; the *m* parameter may be tuned on the basis of observed correlations between causes and effects (for details see Cheng, 1997; Rehder, in press). Finally, it is important to consider the possibility that once a causal model takes hold in one's mental representation of a category, it may be relatively immune to subsequent observations. For example, Rehder and Hastie (2001) systematically examined the effects of causal knowledge and empirical observations, and found that empirical observations had little effect on subsequent classification performance when participants were first provided with causal knowledge.[5] Moreover, Rehder (1999) has demonstrated that causal-model theory yielded better fits to those data sets than the traditional similarity-based models despite the fact that participants were provided with empirical observations in addition to causal knowledge. These findings replicate past results that show that when both prior knowledge and empirical observations are available, performance is often dominated by the knowledge (Chapman & Chapman, 1967, 1969; Murphy & Wisniewski, 1989; Wisniewski, 1995).

Finally, it is worthwhile comparing the classification performance one expects with a causal model versus with the exemplars generated by the model (but without the model itself). Given that the observations are generated from the model, one might predict that classification-by-model and classification-by-observation should converge. However, classification-by-observation is also likely to reflect similarity relations among exemplars. For example, Rehder and Hastie (2001) presented participants with a sample of exemplars generated from a common-cause model (but not the model itself) which did not include exemplar 0111 (common cause absent, all effects present), because 0111 is very unlikely to be generated by that model. Yet, on a subsequent classification test participants granted 0111 a moderately high category membership rating, presumably because 0111 was very similar to the frequently-observed category prototype 1111. That is, classifying on the basis of similarity has the potential to blur theoretically-important distinctions (such as the one between 0111 and 1111). As a result, in many cases classification-by-observation may not be equivalent to classification on the basis of explicit knowledge of the causal rules that generate the observations.

7.3. Causal knowledge and the importance of individual features

Although the emphasis in this article has been on how causal knowledge changes the acceptability of combinations of features to category membership, an equally important question was how that knowledge changes the importance of features considered individually. Earlier research by Ahn and her colleagues (Ahn, 1998; Ahn et al., 2000) suggested that features would be more heavily weighted to the extent they were more causal, that is, more deeply embedded in a network of causal relationships. Similarly, Sloman et al. (1998) have proposed

that features that are "depended on" are less "mutable" (and hence more important to category membership) than those that are not, where a cause–effect link is an example of a dependency relation in which the effect depends on the cause. However, these proposals do not account for the fact that although the common-cause feature was weighed more heavily than its effects, the common-effect was weighed more heavily than its causes.

These findings replicate those of Rehder and Hastie (2001) who found that both common cause and common effect features were weighed more heavily than other features under a wide variety of conditions. These results led Rehder and Hastie to suggest that features are more heavily weighed to the extent they are involved in many causal relationships regardless of whether they play the role of causes and effects. a proposal which may be referred to as the *relational centrality hypothesis*. However, this proposal is also deficient in not accounting for Ahn's and Sloman's finding that the initial cause in a causal chain is weighed most heavily (it predicts that features in the middle of a chain should be weighed most heavily because they are involved in more relationships). It also does not account for the importance of interactions between features, including the asymmetries found with common-cause and common-effect networks which has been the main focus of this article.

Another possibility would be to consider a hybrid account in which features are more heavily weighed both to the extent they are more causal (the causal status hypothesis) and to the extent they are involved in many relationships (the relational centrality hypothesis). However, there are three aspects of the current results that this hybrid account fails to account for. First, it predicts that a common cause feature in a common cause network should be weighed more heavily than a common effect feature in the common effect network, because whereas they are both are involved in the same number of causal relationships (three), the common cause is more "causal" (because it has effects and the common effect does not). Contrary to this prediction, the regression weight associated with the common cause in Fig. 4 (8.2) was *less* than that associated with the common effect (9.0) (albeit not significantly less). Likewise, it predicts that the causes in a common effect network should be weighed more heavily than the effects in a common cause network, because whereas all are involved in the same number of causal relationships (one), the causes are more "causal" than the effects. In fact, the average regression weight associated with the common effect's causes (2.5) was not significantly greater than that associated with the common cause's effects (2.0) ($p > .20$). Finally, this hybrid also does not account for the presence of interactions between features.

Quantitative model fitting revealed that causal-model theory is able to account for the increased weight associated with features that are either common causes or common effects (also with the entire profile of inter-feature interactions shown in Fig. 4). However, because of causal-model theory's asymmetrical view of causal relationships it must have two different accounts for why common causes and common effects dominate (Ahn & Kim, 2001). On the one hand, causal-model theory explains the importance of the common effect feature as a natural consequence of its position in a causal network: because it has many causes it is likely to be generated, and because it is likely to generated it will be heavily weighted in classification judgments. The claim that features increase in importance to the extent they have many causes is a general prediction of causal-model theory that holds for a variety of causal networks (Rehder, in press).[6] On the other hand, because causes do not become more likely as a natural consequence of their having many effects (because giving a cause additional effects does not

make its occurrence more likely) causal-model theory accommodated (but didn't explain) the common cause's importance by adjusting a free parameter (the *c* parameter) that represents the likelihood the common cause was present. To explain the finding that causes sometimes receive greater weight in judgments of category membership, Rehder (in press) has suggested that people often reason with a more complex causal model than the one with which they were provided by experimenters. For example, it has been suggested that people often view categories as being organized around underlying properties or characteristics (sometimes referred to as an *essence*) that are shared by all category members and by members of no other categories, and that essential features cause or generate perceptual features (Gelman et al., 1994; Gelman & Wellman, 1991; Keil, 1989; McNamara & Sternberg, 1983; Medin & Ortony, 1989; Rehder & Hastie, 2001; Rips, 1989). In fact, Rehder (in press) has shown that causal-model theory predicts, and that participants exhibit, a causal status effect with categories that are described as possessing an essential feature and observable features linked in a causal chain. These results can be illustrated with the disease example presented earlier. If one is told that a disease D causes symptom X which causes symptom Y which causes Z, symptom X will be treated as more diagnostic of D than Y (and Y more diagnostic than Z). Causal-model theory predicts this result because D generates X with greater reliability than it does Y (and Y with greater reliability than Z). These results support the claim that the causal status effect observed in causal chains (Ahn, 1998; Ahn et al., 2000; Sloman et al., 1998), and in the current common-cause condition, arises because the features that occupy earlier positions in a causal network are themselves thought to be generated by the unobservable causes that categorizers treat as defining of category membership (Rehder, in press).

8. Conclusion

This article has proposed that the causal relations that link category features are represented in terms of asymmetric and probabilistic causal mechanisms, and that category membership is evaluated on the basis of whether objects were likely to have been generated by those mechanisms. Experimental results confirmed that exemplars were rated as good category members when their features manifested the expectations that causal knowledge induces. The formalization offered by causal-model theory enabled quantitative fits to empirical data, produced interpretable parameter estimates, and supported rigorous tests against competing models. In particular, causal-model theory was shown to achieve a superior fit than extensions to traditional simulate-based models that represent causal knowledge either as higher-order relational features or "prior exemplars" stored in memory.

Notes

1. The common-cause and common-effect models may be defined such that each cause feature has its own independent *c* parameter, each effect feature has its own independent *b* parameter, and each causal link has its own independent *m* parameter. However, initial model fitting results for the data set reported in this article indicated no significant dif-

ferences in either the common-cause condition or the common-effect condition between the parameter estimates for the independent *c*'s, *b*'s, and *m*'s, and so a single *c*, *m*, and *b* parameter was used in each model.

2. Average RMSD is the RMSD averaged over the 36 participants in each condition, where RMSD = SQRT(SSE/($N - P$)), SSE = sum of squared error for a participant, N = the number of data points fit, and P = the number of parameters per model. $N = 16$ and $P = 4$ for all three models presented in Table 5.
3. This difference in the pattern of correlations can be illustrated with the disease example presented earlier. When only a certain subpopulation has a disease (a varying-cause common-cause schema) then the disease's symptoms will be correlated. In contrast, when all members of a population have the disease (a fixed-cause common-cause schema), the symptoms are no longer correlated. This is because the presence of one symptom does not increase the probability of the presence of the disease (which is already present with probability 1). Hence, the probability that other symptoms are present also does not increase.
4. In the Bayes' net literature the common-effect schema shown in Fig. 8a is sometimes referred to as "fuzzy-or" network whereas the common-effect schema shown in Fig. 8b is sometimes referred to as a "fuzzy-and" networks.
5. Although Rehder and Hastie (2001) found that categorizers were sensitive to empirical feature *frequencies* in the presence of causal knowledge (Spalding & Murphy, 1999; Wisniewski, 1995). However, there was little evidence that categorizers were sensitive to the empirical feature *correlations* they observed.
6. Although note that an effect feature will become more probable with of additional causes only when it doesn't already appear with probability 1 (Rehder, in press, Eq. (3)).

Acknowledgment

I thank Woo-Kyoung Ahn and three anonymous reviewers for their comments on previous versions of this manuscript.

Appendix A. Configural-Features Prototype Model

The Configural-Features Prototype Model assumes that the four-dimensional stimulus space used in this article is expanded to include three dimensions that encode whether causal relationships are confirmed or violated. That is, in the common-cause condition each to-be-classified exemplar has three additional dimensions indicating whether the causal relationships $F_1 \rightarrow F_2$, $F_1 \rightarrow F_3$, and $F_3 \rightarrow F_4$ are confirmed or violated. Likewise, in the common-effect condition each exemplar has additional dimensions indicating whether the causal relationships $F_1 \rightarrow F_4$, $F_2 \rightarrow F_4$, and $F_3 \rightarrow F_4$ are confirmed or violated. In addition, the category prototype P is represented as 1111111, where the presence of the feature on the fifth, sixth, and seventh dimension indicates that the causal relationships are confirmed in the prototype. 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,

$$\text{Rating (E)} = K\text{Sim(E, P)} = K\left(\prod_{i=1\ldots7} S_i\right)$$

where $S_i = 1$ if $\text{E}_i = \text{P}_i$ otherwise $S_i = s_i$ where $0 \leq s_i \leq 1$. 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 parameters s_5, s_6, and s_7 are the weights associated with violating causal relationships.

The Configural-Features Prototype Model was fit to the data from each participant. Initial model fitting results in the common-cause condition revealed that the estimates for the weight parameters of the three effect features (i.e., s_2, s_3, and s_4) did not differ significantly from one another, and that the estimates of the three parameters associated with violating causal relationships (i.e., s_5, s_6, and s_7) also did not differ significantly from one another. As a result, s_2, s_3, and s_4 were combined into a single parameter s_e representing the weight associated with all three effect features, and s_5, s_6, and s_7 were combined into a single parameter s_v representing the weight associated with violating causal relationships (and s_1 was renamed s_c, as the weight associated with the single cause feature). Likewise, initial model fitting results in the common-effect condition revealed that the estimates associated with the three cause features (i.e., s_1, s_2, and s_3) did not differ significantly from one another, and that the three parameters associated with violating causal relationships (s_5, s_6, and s_7) did not differ significantly from one another. Thus, s_1, s_2, and s_3 were replaced with the single parameter s_c and s_5, s_6, and s_7 were replaced with the single parameter s_v (and s_4 was renamed s_e as the weight associated with the single effect feature). As a result, model fitting in both the common-cause and common-effect conditions each involved four parameters: K, s_c, s_e, and s_v.

Appendix B. Exemplar-Fragments Model

The Exemplar-Fragments Model assumes that categorizers store in memory a category prototype and also exemplars that represent the causal knowledge (Heit, 1994). The partial exemplars (or exemplar "fragments") that represent the common-cause causal relationships are 11*xx* and 00*xx* (representing the expected correlation between features F_1 and F_2 produced by the $\text{F}_1 \rightarrow \text{F}_2$ causal link), 1*x*1*x* and 0*x*0*x* (representing the $\text{F}_1 \rightarrow \text{F}_3$ link), and 1*xx*1 and 0*xx*0 (representing the $\text{F}_1 \rightarrow \text{F}_4$ link). For the common-effect condition the exemplars are 1*xx*1 and 0*xx*0 (for link $\text{F}_1 \rightarrow \text{F}_4$), *x*1*x*1 and *x*0*x*0 (for link $\text{F}_2 \rightarrow \text{F}_4$), and *xx*11 and *xx*00 (for link $\text{F}_3 \rightarrow \text{F}_4$). According to the Exemplar-Fragments Model, the category membership rating assigned to an exemplar E would be equal to its similarity to the category prototype P (1111) plus the stored exemplars where the relative contribution of these two sources is determined by parameter g. That is,

$$\text{Rating(E)} = K[(1-g)\,\text{Sim (E, P)} + g\,\text{Sim}_{\text{Exemplars}}(\text{E})]$$

where K is a scaling constant and

$$\text{Sim}_{\text{Exemplars}}(\text{E}) = \sum_{\text{e}\in\text{Exemplars}} \text{Sim(e, E)}$$

$$\mathrm{Sim}(x, y) = \prod_{i=1\ldots4} S_i$$

where $S_i = 1$ if $x_i = y_i$ otherwise $S_i = s_i$ where $0 \leq s_i \leq 1$. 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. Parameter g can be interpreted as a measure of how many copies of the exemplars that represent prior knowledge are stored in memory relative to the single copy of the prototype 1111.

When computing similarity between a to-be-classified exemplar and the exemplar fragments representing causal knowledge, I assume an "intersection rule" (Estes, 1994) such that the two dimensions that have missing values in the stored exemplars have no influence. For example, in the common-cause condition the similarity of exemplar 0011 to exemplar fragment 11*xx* would be $(s_1 \cdot s_2)$ and its similarity to fragment 0*xx*0 would $(1 \cdot s_4)$.

The Exemplar-Fragments Model was fit to the data from each participant. As was the case for the Configural-Features Prototype Model, initial model fitting in the common-cause condition revealed that the estimates associated with the three effect features (i.e., s_2, s_3, and s_4) did not differ significantly from one another, and so s_2, s_3, and s_4 were replaced with the single weight parameter s_e, and s_1 was renamed s_c. Likewise, initial model fitting results in the common-effect condition revealed that the estimates associated with the three cause features (i.e., s_1, s_2, and s_3) did not differ significantly from one another, and so were replaced with the single salience parameter s_c and s_4 was renamed s_e. As a result, model fitting in both the common-cause and common-effect conditions each involved four parameters: K, s_c, s_e, and g.

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., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, *54*, 299–352.

Ahn, W., & Kim, N. S. (Eds.). (2001). *The causal status effect in categorization: An overview*. San Diego, CA: Academic Press.

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.

Ahn, A. H., & Lassaline, M. E. (1995). Causal structure in categorization. *Proceedings of the seventeenth annual conference of the cognitive science* (pp. 521–526).

Ahn, W., March, J. K., Luhmann, C. C., & Lee, K. (2002). Effect of theory based correlations on typicality judgements. *Memory & Cognition*, *30*.

Bartlett, F. C. (1932). *Remembering*. Cambridge, England: Cambridge University Press.

Bransford, J. D., Barclay, J. R., & Franks, J. J. (1972). Sentence memory: A constructive versus interpretive approach. *Cognitive Psychology*, *3*, 193–209.

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.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*, 55–81.

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.
Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition*, *40*, 83–120.
Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*, 365–382.
Chi, M. (1993). Conceptual change within and across ontological categories: Examples form learning and discovery in science. In R. Giere (Ed.), *Cognitive models of science: Minnesota studies in the philosophy of science*. Minnesota: University of Minnesota Press.
Chi, M. T. H., Bassok, M., Lewis, M., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*, 145–182.
Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*, 121–152.
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., & Wellman, H. M. (1991). Insides and essences: Early understandings of the nonobvious. *Cognition*, *38*, 213–244.
Gentner, D. (1983). Structure mapping: A theoretical framework for analogy. *Cognitive Psychology*, *15*, 1–38.
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: Oxford University Press.
Gopnik, A., & Meltzoff, A. N. (1998). *Words, thoughts, and theories*. 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. (1998). Influences of prior knowledge on selective weighting of category members. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 712–731.
Heit, E. (2001). Background knowledge and models of categorization. In U. Hahn & M. Ramscar (Eds.), *Similarity and categorization* (pp. 155–178). Oxford: Oxford University Press.
Heit, E., & Bott, L. (2000). Knowledge selection in category learning. In D. L. Medin (Ed.), *The psychology of learning and motivation* (pp. 163–199). Academic Press.
Jordan, M. I. (Ed.). (1999). *Learning in graphical models*. Cambridge, MA: MIT Press.
Karmiloff-Smith, A., & Inhelder, B. (1975). If you want to get ahead, get a theory. *Cognition*, *3*, 195–212.
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.
Kelly, H. H. (1973). The process of causal attribution. *American Psychologist*, *28*, 107–128.
Kemler Nelson, D. G. (1984). The effect of intention on what concepts are acquired. *Journal of Verbal Learning and Verbal Behavior*, *23*, 734–759.
Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*, 163–182.

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

McNamara, T. P., & Sternberg, R. (1983). Mental models of word meanings. *Journal of Verbal Learning and Verbal Behavior*, *22*, 449–474.

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

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

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

Minsky, M., & Papert, S. (1988). *Perceptrons*. Cambridge: 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. (1993). Theories and concept formation. In I. V. Mechelen, J. Hampton, R. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 173–200). London: Academic Press.

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: Theory and applications* (Vol. 2, 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.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgement*. Englewood Cliffs, NJ: Prentice Hall.

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, UK: Cambridge University Press.

Rehder, B. (1999). A causal model theory of categorization. *Proceedings of the 21st annual meeting of the cognitive science society* (pp. 595–600). Vancouver, British Columbia.

Rehder, B. (in press). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

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., & 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.

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

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.

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

Schustack, M. W., & Sternberg, R. J. (1981). Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General*, *110*, 101–120.

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, MA: Harvard University Press.

Spalding, T. L., & Murphy, G. L. (1999). What is learned in knowledge-related categories? Evidence from typicality and feature frequency judgements. *Memory & Cognition*, *27*, 856–867.

Sperber, D., Premack, D., & Premack, A. J. (Eds.). (1995). *Causal cognition: A multidisciplinary approach*. Oxford: Clarendon Press.

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 53–76.

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

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.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*, 337–375.

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.