Memory & Cognition
2009, 37 (6), 715-729
doi:10.3758/MC.37.6.715

Classification as diagnostic reasoning

Bob Rehder and ShinWoo Kim
New York University, New York, New York

An ongoing goal in the field of categorization has been to determine how objects' features provide evidence of membership in one category versus another. Well-known findings include that feature diagnosticity is a function of how often the feature appears in category members versus nonmembers, their perceptual salience, how features are used in support of inferences, and how observable features are related to other observable features. We tested how diagnosticity is affected by causal relations between observable and unobserved features. Consistent with our view of classification as diagnostic reasoning, we found that observable features are more diagnostic to the extent that they are caused by underlying features that define category membership, because the presence of the latter can be (causally) inferred from the former. Implications of these results for current views of conceptual structure and models of categorization are discussed.

It is generally accepted that people's concepts include not only the features and attributes of the entity being represented, but also the ways in which those features are related to one another. For example, we know that hormones can alter a person's behavior, that chemical structure can affect a substance's hardness, and that processor speed can limit a computer's responsiveness. Relational knowledge like this has been shown to affect what people remember, how people reason, and how people use and learn categories of objects. For example, one sort of relational knowledge—causal relations—has been shown to affect a variety of category-related tasks, including how categories are learned (Waldmann, Holyoak, & Fratianne, 1995), to what extent novel properties are generalized to all category members (Heit, 2000; Rehder, 2007; Rips, 2001), and how missing features are inferred (Rehder & Burnett, 2005). In this article, we consider how the inferences licensed by causal knowledge affect the core judgment involving categories—classification itself.

There is an extensive literature documenting how judgments of category membership are affected by the causal relations that link the features that one observes in objects. The *causal status effect* is the phenomenon in which more causal features (features that appear earlier in a category's causal network of features) are more important to category membership than less causal features. For example, holding other factors (like perceptual salience and cue validity) constant, the feature *has wings* should be more diagnostic of birds than the feature *flies*, because flying is a causal consequence of having wings rather than the other way round (Ahn, 1998; Ahn, Kim, Lassaline, & Dennis, 2000; Kim & Ahn, 2002a; Rehder, 2003b; Sloman, Love, & Ahn, 1998). The *multiple cause effect* shows that the causal status effect can be overturned when an effect feature has multiple causes. For example, flying may become more important than wings if flying has additional causes (e.g., having the right body size relative to wing span; Rehder, 2003b; Rehder & Hastie, 2001; Rehder & Kim, 2006). Finally, the *coherence effect* is the phenomenon in which causal relations make objects exhibiting certain combinations of features better category members—namely, those that manifest the interfeature correlations expected to be generated by causal links (e.g., causes and effects either both present or both absent). For example, although atypical, an ostrich is a coherent bird, because it makes sense in light of the causal relations that link features (its large size prevents flying despite the presence of wings) (Marsh & Ahn, 2006; Rehder, 2003a, 2003b; Rehder & Hastie, 2001; Rehder & Kim, 2006).

However, an important omission of these studies is that the causal links investigated have generally been limited to those between observable features (e.g., wings, body size, flying). For example, in the typical study, participants are taught the relations between a small set of three to six features and are then asked to judge the category membership of items in which the presence or absence of each feature is positively affirmed. However, everyday classification typically involves objects in which information about many, if not most, of their features is unavailable. We classify our colleagues into political parties after snippets of conversation, animals from vantage points in which many features are occluded, cars as Toyotas and Fords without looking under the hood, and people as men and women without conducting medical exams. Furthermore, many important properties (DNA, chemical structure) are unobservable without special equipment and training. Indeed, observed features are often just the tip of the iceberg, sitting on top of the rich network of hidden attributes, structures, and processes associated with many kinds.

Of course, the very fact that such features are unobserved provides good reason to think that they play no role

B. Rehder, bob.rehder@nyu.edu

in everyday acts of classification. The brain might simply make do with whatever features are perceptually available, using the evidence that they provide to assign an object to its most likely category. Counter to this view, however, we will argue that unobserved features can play a role in classification, albeit an indirect one, by virtue of the causal relations that link them to the observable ones. For example, wings may be important for identifying birds not because they enable flying, but rather because morphological features like wings are diagnostic of internal structures and processes that are unique to birds. This article tests the hypothesis that classifiers can engage in a kind of two-step inferential process in which they first reason backward from observable features to the unobserved properties or structures and then from those underlying properties to category membership. Following this view, one diagnoses category membership in the same way that one diagnoses the presence of a disease from the presence of the symptoms that it causes. We will refer to this proposal as the *classification as diagnostic reasoning* view.

There is good reason to suspect that the presence of underlying structures implied by observable features can affect classification, because research has shown that those structures are uniquely associated with category membership. For example, in Keil's (1989) well-known transformation experiments, second graders were told about doctors who dyed a raccoon's fur black, bleached a white stripe down its back, and put a sac of odor in its body. The children judged that the transformed animal was still a raccoon, despite its now looking like a skunk, reflecting the importance of the animal's internal versus external properties (see Rips, 1989, and Hampton, Estes, & Simmons, 2007, for related findings with adults). In fact, evidence that the internal structure of animals is more important than the outside has been found with children as young as 3 years old (Gelman & Wellman, 1991; also see Diesendruck, 2001, Gelman, 2003, and Hirschfeld, 1996). But although these studies established the strong relationship between unobserved internal properties and category membership, it may be that classifiers make use of this knowledge only in unusual hypothetical situations involving transformations. Whether they also do so in the acts of classification that people perform every day is an open question.

In this article, we ask whether underlying features can affect categorization even in situations not involving transformations by virtue of their causal links to observable features (Medin & Ortony, 1989). In the following experiments, adults were taught a pair of novel categories and then asked to choose which category a particular object was more likely to belong to. Novel categories were used to control which features were unobservable and which observable features were causally generated by the unobservable ones. For example, some participants learned about two species of ants, Kehoe ants and Argentine ants (Table 1). For Kehoe ants, the underlying feature was *blood high in iron sulfate* and the two observable features might be *hyperactive immune system* and *thick blood*. For Argentine ants, the underlying feature was *blood high in metallic sodium* and the two observable features might be *fast digestion* and *short life span*. Features were *observable* because they were displayed, in various combinations, by the objects that the participants subsequently classified, whereas no explicit information regarding the presence of the underlying features in those objects was provided. Importantly, one or more of the observable features were described as being causally generated by the underlying one. For Kehoe ants, the participants might be told that "blood high in iron sulfate causes a hyperactive immune system" and about the causal mechanism linking those two features: "The iron sulfate molecules are detected as foreign by the immune system, and the immune system is highly active as a result." Examples of causal relations relating underlying and observable features are presented in Table 1 for Kehoe and Argentine ants. Five other pairs of categories besides ants were tested.

Our central hypothesis is that observable features become more diagnostic of category membership by virtue of the causal relations that link them to an underlying feature, because from observable features (e.g., hyperactive immune system) one can reason causally to an underlying feature (e.g., blood high in iron sulfate) and then to category membership (e.g., Kehoe ants). To instantiate the strong relationship between underlying features and category membership established by Keil (1989), Rips (1989), Gelman (2003), and others, the diagnosticity of those properties was maximized by stipulating them to be *defining features*—that is, to covary perfectly with their category. For example, *blood high in iron sulfate* was stipulated as defining of Kehoe ants by stating that it was present in all Kehoe ants and in no other species of ants. However, it is not our view that diagnostic reasoning in the service of classification is limited to defin-

Table 1
Example of Experimental Materials

Category	Underlying Feature	Observable Features	Causal Relations
Kehoe ants	Blood high in iron sulfate	Hyperactive immune system	Blood high in iron sulfate causes a hyperactive immune system. The iron sulfate molecules are detected as foreign by the immune system, and the immune system is highly active as a result.
		Thick blood	Blood high in iron sulfate causes thick blood. Iron sulfate provides the extra iron that the ant uses to produce extra red blood cells. The extra red blood cells thicken the blood.
Argentine ants	Blood high in metallic sodium	Fast digestion	Blood high in metallic sodium causes fast digestion. Because metallic sodium is a digestive enzyme that facilitates nutrition extraction, high levels of metallic sodium result in fast digestion.
		Short life span	Blood high in metallic sodium causes a short life span. Because metallic sodium gradually corrodes the valves in the ant's heart, its life span is shorter.

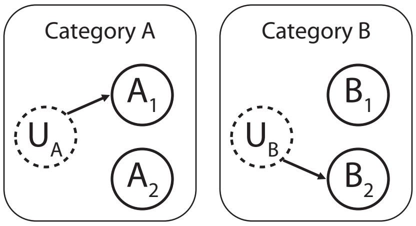

Figure 1. Category structures tested in Experiment 1.

ing features. Rather, any underlying feature inferred from observable ones can affect classification so long as it provides reasonably strong evidence of category membership.

The following experiments provide direct tests of the claims of the classification as diagnostic reasoning view. Experiment 1 confirms that a feature is indeed more diagnostic when it is perceived as being caused by its category's underlying properties. Experiments 2 and 3 rule out an alternative interpretation that this effect is due to the additional salience that a feature accrues by being involved in causal relations. Experiments 3 and 4 directly demonstrate that the increase in diagnosticity is due to the reasoning from observable to unobservable features. Finally, Experiment 5 demonstrates that this reasoning is causal in nature by exhibiting the asymmetries that are inherent in causal relations.

Besides being important in their own right, the present experiments also provide tests of a number of competing accounts of how causal knowledge affects classifications. For example, according to Sloman et al.'s (1998) dependency model, features vary in diagnosticity as a function of the number of dependents (i.e., effects) that they have. According to Rehder and Murphy's (2003) knowledge–resonance (KRES) model, unobserved features can affect classification via a constraint satisfaction process in which they are activated by observed features. Finally, Rehder's (2003a, 2003b; Rehder & Kim, 2006) generative model predicts that features are diagnostic of category membership to the extent that are generated or produced by a category's causal model. In the General Discussion section, we will consider the implications of our results for these models after we present our empirical findings.

EXPERIMENT 1

To conduct an initial test of the proposal that features are more diagnostic when they are causally related to an underlying feature, the participants in Experiment 1 were taught the two novel categories shown in Figure 1. Category A had three features, one underlying feature (U_A) and two observable features (A_1 and A_2). The first observable feature (A_1) was described as being caused by U_A, but the second (A_2) was not. Likewise, Category B had one underlying feature (U_B) that caused the second observable feature (B_2) but not the first (B_1). For example, of the participants who learned the two species of ants, half were told that Kehoe ants' underlying feature (*blood high in iron sulfate*) caused its first observable feature (*hyperactive immune system*) but not its second (*thick blood*), and that Argentine ants' underlying feature (*blood high in metallic sodium*) caused its second observable feature (*short life span*) but not its first (*fast digestion*). As was previously mentioned, U_A and U_B were defining because they were described as occurring in all members of their respective category and in no nonmembers. Observable features were associated with their category by stating that they occurred in 75% of the category members. To eliminate the possibility that any effects are due to the particular features and causal relationships involved, the assignment of Kehoe ants and Argentine ants to the roles of Category A or B in Figure 1 was reversed for the other half of the participants. Five other category pairs besides ants were tested.

After learning about the two categories, the participants performed classification tests in which they were shown two features, one from each category, and asked which category the object belonged to. For example, a test item might have features A_1 and B_1, which we predict will be classified as belonging to Category A, because, from A_1, one can reason to U_A via the causal link that connects them, but one cannot so reason from B_1 to U_B. For a similar reason, an item with features A_2 and B_2 should be classified as belonging to Category B. We refer to those test items with features whose presence is affirmed as *positive items*. We also presented *negative items* with features that were stipulated as absent. For example, an ant might have a normal rather than a hyperactive immune system (which we denote as $\sim A_1$) and a normal rather than fast digestion ($\sim B_1$). In this case, we predict that the item will be classified as belonging to Category B, because, from $\sim A_1$, one can reason that U_A is likely to be absent. The test items and predictions are listed in Table 2.

Method

Materials. Six pairs of categories were tested: Kehoe and Argentine ants; Lake Victoria and Madagascar river shrimp; myastars and terastars; meteoric sodium carbonate and terrestrial sodium carbonate; Romanian Rogos and Bulgarian Bentos (types of automobiles); and Neptune personal computers and Martian notebook computers. Note that, for generality, these pairs include both artifacts and natural kinds that were both biological and nonliving. Each category had three features, where one was unobserved and two were observable, and the observable features could be causally related to the underlying one (see Table 1 for an example). The features and causal relationships for all categories are available, on request, from the authors.

Design. Each participant learned one of the six pairs of categories. In addition, there were two between-subjects counterbalancing factors. First, which members of a category pair played the roles of Category A and B was balanced. For example, of those participants assigned to the two ant categories, Categories A and B were instantiated by Kehoe ants and Argentine ants, respectively, for half of the partici-

Table 2
Test Items and Results From Experiment 1

		Predicted	*P*(Predicted Category)		Signed Confidence	
Test Item Type	Test Item	Category	*M*	*SE*	*M*	*SE*
Positive	A_1B_1	A	$.84^{**}$	.05	$+50^{**}$	5
	A_2B_2	B				
Negative	$\sim A_1 \sim B_1$	B	$.68^{**}$	.06	$+21^{*}$	6
	$\sim A_2 \sim B_2$	A				

Note—Proportion of predicted category responses and signed confidence ratings were tested against .50 and 0, respectively. $^{*}p <$ .05. $^{**}p < .01$.

pants and this assignment was reversed for the other half. As a result of this balancing, the results of the classification test averaged over the effects of the physical features involved (e.g., *hyperactive immune system* vs. *fast digestion*). Second, half of the participants learned Category A then Category B and vice versa for the other half.

Participants. Twenty-four New York University undergraduates participated for course credit. They were randomly assigned in equal numbers to one of the six category pairs, to one of the two category orders, and to one of the two assignments of the two physical categories to roles of Category A or B.

Procedure. The experiment was conducted by computer with intermittent spoken instructions. The participants first learned the two categories, one after the other. For each, they studied several screens of information about the category, including a cover story, information on features and causal relations, and a summary diagram of causal relations much like Figure 1 (with category and feature descriptions substituted for the abstract labels). When ready, the participants took a multiple-choice test of 11 questions. While taking the test, the participants were free to return to the study screens; however, doing this obligated the participant to retake the test. The only way to pass the test and proceed to subsequent phases was to complete it without errors and without returning to the initial study screens for help.

After learning the two categories, the participants proceeded to the classification test. During the test, they were allowed to refer to printed diagrams of the two categories' features and causal relations. The test items consisted of pairs of observable features, one from each category (e.g., A_1B_1). Each test item was presented as two lines of text (e.g., *hyperactive immune system* and *fast digestion*), one below the other in random order. After classifying each item into one of the two categories, the participants also provided confidence ratings by positioning a slider on a scale whose left and right ends were labeled *very uncertain* and *very certain*. The slider could be set to 21 distinct positions, and the responses were scaled to range from 0 to 100. The classification test consisted of two blocks of four test items (A_1B_1, A_2B_2, $\sim A_1 \sim B_1$, and $\sim A_2 \sim B_2$). The items were presented in a different random order within each block. The experimental sessions lasted approximately 30 min.

Results

Initial analyses revealed no effects of category pair, the order of category presentation, or the assignment of pairs to Category A or B, and the results are therefore presented in Table 2 collapsed over these factors. Note that because of counterbalancing, the two positive items (A_1B_1, A_2B_2) are logically identical to one another, as are the two negative items ($\sim A_1 \sim B_1$, $\sim A_2 \sim B_2$), and so the results are collapsed over these items as well. Table 2 presents the proportion of categorization decisions consistent with the predicted category (e.g., Category A for A_1B_1, Category B for A_2B_2, etc.). It also presents the mean signed confidence ratings for each item type. For each trial, the signed confidence rating was set equal to the participants' confidence ratings (0–100) if they classified the item into the predicted category, and to the negated confidence rating if the item was classified into the other category. Signed confidence ratings thus provide a secondary measure of the participants' categorization preferences, with positive numbers indicating a preference for the predicted category and negative numbers a preference for the other category (and 0 indicating no preference). Finally, Table 2 also includes, for each item type, the results of *t* tests of whether the participants' categorization decisions differed from .5 and whether their signed confidence ratings differed from 0.

Table 2 indicates that, as was predicted, features were more diagnostic of category membership when they were causally related to an underlying feature. Positive items A_1B_1 and A_2B_2 were classified significantly more often (.84) into Categories A and B, respectively, apparently because A_1 implies U_A and B_2 implies U_B. The signed confidence ratings for the positive items (50) were also significantly greater than 0. The predictions of the diagnostic reasoning account were also confirmed in the negative items $\sim A_1 \sim B_1$ and $\sim A_2 \sim B_2$. An item should not be classified into a category (e.g., Category A) whose underlying defining feature can be inferred to be absent (e.g., $\sim A_1$ implies $\sim U_A$). In fact, the choice probability for the negative items (.68) also differed significantly from .5, and the signed confidence ratings (21) were significantly higher than 0, in accord with this prediction.[1]

Discussion

As was predicted, a feature was more diagnostic of category membership when it was causally related to an underlying feature. Moreover, the explicit absence of that feature counted for stronger evidence against category membership. According to the diagnostic reasoning account, these results arose because classifiers reasoned causally from the presence (or absence) of an observed feature to the presence (or absence) of the underlying one and then to category membership.

EXPERIMENT 2

Experiment 1 showed that when a feature is causally related to an underlying feature, it becomes more diagnostic of category membership. Although our interpretation of this result is that classifiers reasoned from observed to unobserved features (and then to category membership), an alternative explanation is that the causally related observable features merely became more salient because of their participation in the causal link (Ahn & Kim, 2001). This may have been the case because, for example, the causally related features were mentioned and asked about more often in the self-paced learning and the multiple-choice test, and this repetition alone may have resulted in those features being treated as more important.

In Experiment 2, we addressed this alternative explanation by testing the categories in Figure 2. In Category A, U_A caused A_1 (as it did in Experiment 1), but in addition, A_1 itself caused A_2. In contrast, B_1 caused B_2 in Category B, but neither observable feature was causally related to U_B. When presented with test item A_1B_1, we of course predict that the item will be classified as a member of Category A, because A_1 implies the presence of U_A. However, the critical test item in Experiment 2 was A_2B_2. According to the diagnostic reasoning account, A_2B_2 should be classified as a member of Category A, because, whereas from A_2, one can infer A_1 and then U_A, B_2 implies B_1 but not U_B (because of the lack of a causal link between B_1 and U_B). In contrast, the alternative salience account predicts that participants should be agnostic regarding A_2B_2, because both A_2 and B_2 are involved in exactly one causal link (and were thus mentioned an equal number of times during the

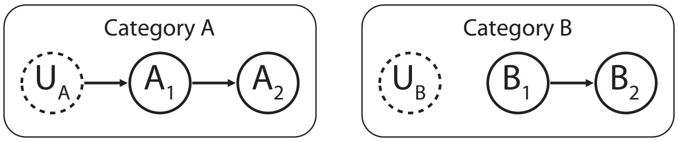

Figure 2. Category structures tested in Experiment 2.

self-paced learning). Analogously, the diagnostic reasoning account predicts that both negative items should be classified as members of Category B, whereas the salience account predicts no preference for item ~A_2~B_2. The four test items and predictions of the diagnostic reasoning account are listed in Table 3.

Method

The materials and procedure in Experiment 2 were identical to those in Experiment 1, except for the different causal relationships required by the networks in Figure 2. Twenty-four New York University undergraduates participated for course credit. They were randomly assigned in equal numbers to one of the six category pairs, to one of the two category orders, and to one of the two assignments of physical to logical categories.

Results

Initial analyses again revealed no effect of category pair or of the two counterbalancing factors, and, therefore, collapsed results are presented in Table 3. The proportion of categorization choices and signed confidence ratings were computed in the same ways as were those in Experiment 1.

Table 3 shows that all choice proportions were significantly higher than .5, supporting the diagnostic reasoning account. Consistent with the results of Experiment 1, item A_1B_1 was typically classified as a member of Category A (because A_1 implies the presence of U_A), and its negative counterpart, ~A_1~B_1, was classified more often as a member of Category B (because ~A_1 implies ~U_A). The critical test items, however, were A_2B_2 and ~A_2~B_2, because, whereas the alternative salience account predicts no preference, the diagnostic reasoning account predicts that they should be classified as members of Category A and Category B, respectively. In fact, A_2B_2 was classified more often as a member of Category A, indicating that A_2 provided stronger evidence for Category A than B_2 did for Category B. Conversely, ~A_2~B_2 was classified more often as a member of Category B. The signed confidence ratings for the two items were also higher than 0.

Another notable aspect of the results was that A_1B_1 was classified more decisively as a member of Category A than was A_2B_2, and ~A_1~B_1 was classified more decisively as a member of Category B than was ~A_2~B_2. Indeed, a 2 × 2 repeated measures ANOVA of the choice proportions in Table 3 with dimension (1 vs. 2) and item type (positive vs. negative) as factors revealed an effect of dimension [$F(1,23) = 6.82$, $MS_e = .086$, $p < .05$] but no effect of item type and no interaction (Fs < 1). Our interpretation of this result is that A_1 provides stronger evidence in favor of Category A than A_2, because one infers U_A from A_1 over one link, whereas the inference from A_2 is over two links, and multilink inferences are less certain when causal relations are probabilistic.

Discussion

Experiment 1 showed that features causally related to an underlying property were more heavily weighed than unrelated features. The results of Experiment 2 rule out the possibility that this effect was merely due to the features' involvement in just any kind of causal relation. Although A_2 and B_2 were both involved in a single causal relationship, only A_2 implied the presence of its respective underlying feature, and, in fact, A_2 served as stronger evidence for Category A than B_2 did for Category B. Apparently, classifiers can reason from observable features to underlying ones, and decide category membership on that basis.

EXPERIMENT 3

Experiment 3 provides another test of the feature salience account by manipulating the reliability of the causal processes that generate observable features. According to the diagnostic reasoning account, classifiers reason backward from observable to unobservable features (and then to category membership). Thus, this reasoning should be made more confidently when those features are related by a more reliable causal process. This prediction was tested in Experiment 3 by use of the category structures seen in Figure 3. Whereas in Experiments 1 and 2, no information about the reliability of the causal processes was provided, in Experiment 3 the reliability of the causal processes between U_A and A_1 and that between U_A and A_2 were described as 90% and 60%, respectively, and those between U_B and B_1 and between U_B and B_2 were described as 60% and 90%, respectively. For example, when Kehoe ants instantiated Category A, the causal link between blood high in iron sulfate and a hyperactive immune system was described with the additional sentence "Whenever a Kehoe ant has blood high in iron sulfate, it will cause that ant to have a hyperactive immune system with probability 90%."

We predicted that the test item A_1B_1 would be classified as a member of Category A, because A_1 is produced more reliably by U_A than B_1 is produced by U_B (and, thus, the

Table 3
Test Items and Results From Experiment 2

		P(Predicted Category)		Signed Confidence	
Test Item	Predicted Category	*M*	*SE*	*M*	*SE*
A_1B_1	A	$.81^{**}$	.06	$+36^{**}$	7
A_2B_2	A	$.67^{*}$	.07	$+16^{\dagger}$	9
~A_1~B_1	B	$.83^{**}$	.06	$+36^{**}$	9
~A_2~B_2	B	$.67^{*}$	.07	$+18^{*}$	8

Note—Proportion of predicted category responses and signed confidence ratings were tested against .50 and 0, respectively. $^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$.

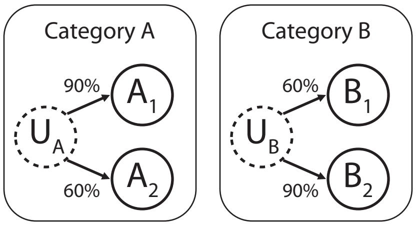

Figure 3. Category structures tested in Experiment 3.

inference from A_1 to U_A is more certain than the inference from B_1 to U_B). Likewise, the test item $\sim A_1 \sim B_1$ should be classified as a member of Category B, because $\sim A_1$ counts as greater evidence against U_A than $\sim B_1$ does against U_B. The four test items and predictions are presented in Table 4. If obtained, these predicted results will confirm the importance of the reliability of causal processes associated with a category and will provide further evidence that the results from Experiment 1 were not merely due to involvement in a causal relationship (e.g., A_1 and B_1 were both involved in one relationship) or to how often they are mentioned in the instructions.

Method

Materials. The causal relationships in Experiment 3 were the same as those in the first two experiments, except for the information about the reliability of the cause mechanisms. In addition, the diagrams of the causal links presented during the initial tutorial and the classification test were annotated with "90%" and "60%" on the appropriate links. Whereas in Experiments 1 and 2, each feature was described as occurring in 75% of category members, in the present experiment they were described as occurring in "most" members of the category, in order to avoid any apparent contradiction between features' base rates and the reliability of the causal links.

Participants. Twenty-four New York University undergraduates participated for course credit. They were randomly assigned in equal numbers to one of the six category pairs, to one of the two category orders, and to one of the two assignments of physical to logical categories.

Procedure. The procedure was identical to those of Experiments 1 and 2.

Results

As in the first two experiments, initial analyses revealed no effect of category pair or the counterbalancing factors, and, thus, the results were collapsed over these factors. In addition, because the two positive (A_1B_1, A_2B_2) and negative ($\sim A_1 \sim B_1$, $\sim A_2 \sim B_2$) items were logically identical, the results were also collapsed over these items.

Table 4 indicates that all choice proportions were consistent with the categories predicted by the diagnostic reasoning account. The positive items, A_1B_1 and A_2B_2, were classified more often into Categories A and B, respectively. Apparently, the greater reliability of the causal processes that connected A_1 with U_A and B_2 with U_B meant that those features were weighed more heavily than features generated with less reliability (A_2 and B_1). This interpretation is also supported by the fact that negative items, $\sim A_1 \sim B_1$ and $\sim A_2 \sim B_2$, were classified more often into Categories B and A, respectively. Signed confidence ratings for all of the test items were significantly greater than 0.

Discussion

According the diagnostic reasoning view, the inferences from observable to underlying features should be more certain when the causal processes that related those features are more reliable. The present results provide support for this claim, because the features that were generated with 90% reliability had more influence on classification than features generated with 60% reliability.

The present findings also augment those of Experiment 2, which demonstrated that features are more heavily weighed not merely because they are involved in some kind of causal relationship. For example, although features A_1 and B_1 were both involved in a causal relationship, it was A_1—the feature that was more reliably generated by its respective underlying property—that had the greater impact on classification.

EXPERIMENT 4

According to our account, classification can involve a two-step process in which one first reasons from observed to unobserved features and then from unobserved features to category membership. In Experiments 2 and 3, we tested the first part of this claim by disrupting the causal linkage between observed and unobserved features (Experiment 2) and by varying the strengths of those links (Experiment 3). In Experiment 4, we tested the second part by varying the degree to which the underlying feature is diagnostic of category membership. The participants were instructed on the two categories shown in Figure 4. In both categories, the two observed features were caused by the underlying one. However, whereas U_A was described as occurring in all Category A members (and in no nonmembers) just as in Experiments 1–3, U_B was described as occurring in only 75% of Category B members (and no statement was made about other categories). That is, U_B is no longer a defining feature.

Our predictions are that, whereas the observable features of both categories provide equal evidence for U_A and U_B, respectively, those of Category A should be more diagnostic, because U_A itself is. For example, the positive test items A_1B_1 and A_2B_2 should be classified as a belonging to Category A, because U_A is more diagnostic than U_B. For the converse reason, the negative test items $\sim A_1 \sim B_1$ and $\sim A_2 \sim B_2$ should be classified as members of Category B. In addition to the two-feature test items tested in Experiments 1–3, the four-feature items $A_1A_2B_1B_2$ and $\sim A_1 \sim A_2 \sim B_1 \sim B_2$ were also tested. The predictions are summarized in Table 5.

Table 4
Test Items and Results From Experiment 3

			P(Predicted Category)		Signed Confidence	
Test Item Type	Test Item	Predicted Category	*M*	*SE*	*M*	*SE*
Positive	A_1B_1 A_2B_2	A B	$.88^{**}$	.05	$+51^{**}$	4
Negative	$\sim A_1 \sim B_1$ $\sim A_2 \sim B_2$	B A	$.96^{**}$	.06	$+57^{*}$	4

Note—Proportion of predicted category responses and signed confidence ratings were tested against .50 and 0, respectively. $^{*}p <$ .05. $^{**}p < .01$.

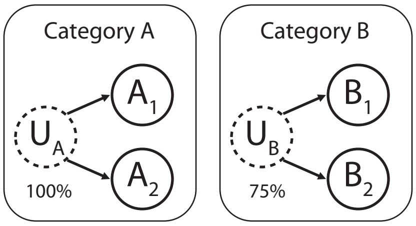

Figure 4. Category structures tested in Experiment 4.

Method

The materials and procedure were the same as those in the first three experiments, except for the 75% base rate for U_B. As in Experiments 1 and 2, no information about the reliability of the causal processes was presented, and observable features were described as occurring in 75% of category members. Twenty-four New York University undergraduates participated for course credit. They were randomly assigned in equal numbers to one of the six category pairs, to one of the two category orders, and to one of the two assignments of physical to logical categories.

Results

Once again, initial analyses revealed no effect of category pair or the counterbalancing factors, and, therefore, the results were collapsed over these factors and also over the two positive (A_1B_1, A_2B_2) and negative (~A_1~B_1, ~A_2~B_2) items. Table 5 indicates that all choice proportions were consistent with the predictions. Both the two- and four-feature positive items were classified more often into Category A, apparently because the evidence in favor of Category A provided by U_A (whose presence is implied by A_1 and A_2) is greater than the evidence in favor of Category B provided by U_B (whose presence is implied by B_1 and B_2). This is the case because, whereas U_A is essential to Category A, U_B is only probabilistically associated with Category B. Conversely, the negative items were judged more often to be in Category B, because the inferred absence of U_A provides stronger evidence against Category A than the absence of U_B provides against Category B. Signed confidence ratings for all test items were also significantly higher than 0.

Discussion

As was predicted, a feature is more diagnostic when its underlying cause appears in all category members (and in no nonmembers) than when it appears just in most category members—that is, when it is defining of category membership. Importantly, this experiment provides the most direct evidence so far that classifiers infer unobservables in the service of categorization, because Categories A and B only differed on properties (U_A and U_B) that were themselves never observed during classification. Apparently, the participants inferred the likely presence (or absence) of these underlying causes and made their judgments on the basis of the cause that was more firmly associated with its respective category.

EXPERIMENT 5

Experiments 1–4 established that people can reason from observed features to underlying ones when deciding category membership. However, we have yet to demonstrate that that reasoning is specifically causal in nature. To accomplish this, the participants in Experiment 5 were taught the two categories shown in Figure 5. In each category, the two observable features were caused by the underlying one. However, unlike in the previous experiments, the participants were given explicit information about the possibility of alternative causes of the observable features; specifically, they were told that one feature had alternative causes, whereas the other had none. For example, when Kehoe ants played the role of Category A, the participants were told that feature A_1 (a hyperactive immune system) had an alternative cause or causes: "Besides blood high in iron sulfate, there are also one or more unknown causes of hyperactive immune system. Because of this, a hyperactive immune system occurs in 50% of ants that don't have blood high in iron sulfate." They would also be told that A_2 (thick blood) had no other causes: "Because there are no other causes of thick blood, ants that don't have blood high in iron sulfate never have thick blood." In the contrast category (e.g., Argentine ants), feature B_2 had alternative causes and B_1 had none.

Our prediction is that features are more diagnostic of their category when they do not have alternative causes. For example, test item A_1B_1 should be classified as a member of Category B, because B_1 provides decisive evidence of U_B (because it has no other causes). In contrast, because it might have been caused by something else, A_1 provides relatively weaker evidence for U_A. That is, inferences from A_1 exhibit a hallmark of causal reasoning known as *discounting*, in which the potential presence of one cause (in this case, A_1's background causes) reduces the likely presence of another (U_A). A_2B_2 should be classified as a member of Category A

Table 5
Test Items and Results From Experiment 4

Test Item Type	Test Item	Predicted Category	*P*(Predicted Category)		Signed Confidence	
			M	*SE*	*M*	*SE*
Positive/two features	A_1B_1 A_2B_2	A A	.68*	.07	+20**	7
Negative/two features	~A_1~B_1 ~A_2~B_2	B B	.66*	.06	+21**	6
Positive/four features	$A_1A_2B_1B_2$	A	.67*	.08	+20*	8
Negative/four features	~A_1~A_2~B_1~B_2	B	.77**	.07	+31**	6

Note—Proportion of predicted category responses and signed confidence ratings were tested against .50 and 0, respectively. $^*p < .05$. $^{**}p < .01$.

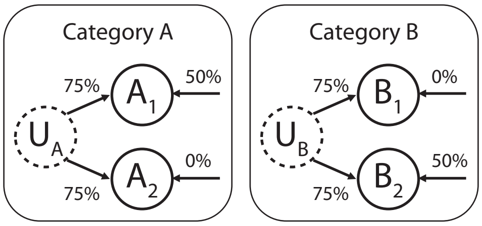

Figure 5. Category structures tested in Experiment 5.

for a similar reason. This prediction reflects causal reasoning, because it exhibits the asymmetries inherent in such reasoning: If the direction of causal links in Figure 5 were reversed, A_1 would provide as much evidence for U_A as B_1 provides for U_B (Kelley, 1973; Morris & Larrick, 1995).

Unlike Experiments 1–4, this experiment yielded no definite predictions for the negative test items, because performance on those items depended on participants' assumption about the alternative causes of A_1 and B_2. On one hand, if the participants assumed that these alternative causes were uncorrelated with either Category A or Category B, they should be at chance on both negative test items. For example, the absence of A_1 implies the absence of U_A as strongly as the absence of B_1 implies the absence of U_B, and, thus, the participants should be at chance for test item $\sim A_1 \sim B_1$. On the other hand, if the participants assumed that the alternative cause of A_1 was in fact Category B (e.g., U_B), $\sim A_1 \sim B_1$ should be classified as a member of Category A, because, whereas the evidence against U_A consists of $\sim A_1$ alone, the evidence against U_B consists of both $\sim A_1$ and $\sim B_1$. (Item $\sim A_2 \sim B_2$ should be classified as a member of Category B for a similar reason.) Our predictions for Experiment 5 are summarized in Table 6.[2]

Method

The materials and procedure were the same as those in Experiments 1–4. The observed features were described as occurring in "most" of their respective category members. Twenty-four New York University undergraduates were randomly assigned in equal numbers to one of the six category pairs, to one of the two category orders, and to one of the two assignments of physical to logical categories.

Results

There was no effect of category pair or the counterbalancing factors, and, therefore, the results were collapsed over these factors in Table 6. The table indicates that the choice proportions of the positive test items were consistent with our predictions. Item A_1B_1 was classified as a member of Category B, because B_1 provided decisive evidence of U_B, and item A_2B_2 was classified as a member of Category A, because A_2 provided decisive evidence of U_A. The participants were at chance for the negative items, suggesting that the alternative causes of A_1 and B_2 were viewed as uncorrelated with either Category A or B.

Discussion

As was predicted, in this experiment, the participants exhibited causal reasoning by discounting—that is, by inferring underlying features less strongly when they were described as having alternative causes.[3] Note that these results also rule out the alternative hypothesis, first raised in Experiment 1, that features become more diagnostic merely when they are involved in more causal relations, because, in this experiment, features A_1 and B_2 were *less* diagnostic even though they were involved in *more* causal relations.

Discounting has been observed in other situations in which one uses causal knowledge to predict the presence of an unobserved category feature. For example, Rehder and Burnett (2005) found that the presence of a cause of a common effect was judged as less certain when other causes were present. The present results extend those findings to situations in which the inference is to a defining feature and in which the participants are asked to judge category membership. In addition, of course, discounting has been observed in the causal attribution literature in social psychology (see McClure, 1998, for a review).

Additional experiments in our lab have established the role of causal reasoning in classification by demonstrating the asymmetries inherent in such reasoning. For example, we have instructed participants on categories whose features were either caused by an underlying property (as in Experiments 1–5) or were the cause of that property. Two-feature test items consisting of a cause feature from Category A and an effect feature from Category B were classified as members of Category A, indicating that the inference to an underlying feature was stronger in the forward (cause-to-effect) than in the backward (effect-to-cause) direction, consistent with the well-known result that people reason more confidently from causes to effects than they do vice versa (Tversky & Kahneman, 1980). In the General Discussion section, we will review additional evidence of causal processing in a variety of category-based judgments.

GENERAL DISCUSSION

In this article, we have asked how classification is affected when underlying properties are causally linked to observable ones. In the following sections, we will discuss our findings regarding the role of causally generated features in categorization and the ability of current models of causal-based categorization to account for those results.

The Diagnosticity of Causally Generated Features

Since the inception of the probabilistic view of categorization, an ongoing research goal has been to identify how

Table 6
Test Items and Results From Experiment 5

			P(Predicted Category)		Signed Confidence	
Test Item Type	Test Item	Predicted Category	*M*	*SE*	*M*	*SE*
Positive	A_1B_1	B	.73**	.08	+42**	1
	A_2B_2	A				
Negative	$\sim A_1 \sim B_1$	–	.46[a]	.07	−10[a]	7
	$\sim A_2 \sim B_2$	–	.48[a]	.09	−1[a]	10

Note—Proportion of predicted category responses and signed confidence ratings were tested against .50 and 0, respectively. [a]Because there is no predicted category for the negative items, choice proportions and signed confidence ratings for these items are reported relative to Category A. $^{**}p < .01$.

classifiers use an object's features to determine its category membership. Well-known past findings include that feature diagnosticity is affected by empirical information—that is, how often features appear in category members and in nonmembers (Hampton, 1979; Medin & Schaffer, 1978; Rosch & Mervis, 1975), how features are used in category-based inferences (Ross, 1996, 1997, 1999), and perceptual salience (Sloman et al., 1998). More recent studies have documented a number of effects indicating how diagnosticity is affected by causal relations between observable features (e.g., Ahn et al., 2000; Rehder & Hastie, 2001). The contribution of this research is that it establishes how features' diagnosticity is also determined by the causal relations linking them to underlying features.

The importance of causally generated features would seem to characterize our understanding of many real-world categories. For example, although most adults believe that gender is defined by underlying biological properties (e.g., chromosomes), to perform the everyday act of identifying individuals as men and women, they must rely on observable properties. But we recognize that not all observable properties are equal, because, whereas some are causally linked to underlying biology (e.g., body shape, voice pitch), others are more determined by cultural conventions (e.g., hair length, clothing style). Thus, even though both sorts of features contribute to categorizations, we recognize that the former are—all else being equal—more reliable cues than the latter.

Our participants' sensitivity to the specific properties of the causal links relating underlying and observable features indicated that they were engaged in a process of causal inference in which they reasoned from observable features to unobservable ones and then to category membership. In Experiment 3, we showed that observable features were more diagnostic for stronger than for weaker causal links. This result is consistent with classifiers' reasoning from observable to unobservable features, given that such inferences will be made with greater certainty for more reliable causal processes. Similarly, in Experiment 2, we found that those inferences were less certain when another feature intervened between the observable and unobservable features, consistent with the idea that inferences over many variables will be viewed as less reliable than those over fewer. Finally, Experiment 5 showed that features are more diagnostic of their category when they have no alternative causes, consistent with a discounting effect in which the presence of one potential cause is made less likely by the presence of alternative causes.

The presence of diagnostic reasoning in classification is important, because it suggests that causal knowledge contributes to the many documented cases of "fuzziness" in natural categories (Hampton, 1979; McCloskey & Glucksberg, 1978). According to the traditional view, these effects arise because features are observed to be only probabilistically associated with their category. But according to our account, it also arises from the inferential uncertainty that results from classifiers' beliefs about features' involvement in causal relations: The evidence that those features provide for underlying properties and structures varies depending on the strength of the causal links (and the probability that those features were caused by the underlying properties of other categories). Again, this sort of understanding appears to apply naturally to many real-world categories. Because we know that flying is generated by the causal mechanisms of birds with less than perfect reliability, not all birds fly (e.g., ostriches, penguins); because we know that flying is generated by the causal mechanisms of other categories, not all nonbirds fail to fly (e.g., mosquitoes, airplanes). In this manner, the probabilistic information provided by causal knowledge combines with the probabilistic information that we gather from firsthand observations to determine the degree of evidence that features provide for category membership (McNorgan, Kotack, Meehan, & McRae, 2007).

Our diagnostic reasoning account is related to the view of conceptual structure known as *psychological essentialism*. On this account, people view categories as having underlying properties and structures that are *essential*—that is, that make an object the kind of thing that it is (Gelman, 2003; Medin & Ortony, 1989). Like the underlying features in Experiments 1–5, essences are defining (present in all category members and in no nonmembers) and they constrain, or generate, the features of objects that can be observed. However, unlike our underlying features, truly essential features are also viewed as immutable and having innate origins; indeed, essential properties are ones that are present in all category members that *could* exist (Gelman, 2003). And, although adults may have concrete beliefs about essences (e.g., DNA functions as the essence for biological kinds for many Western-educated adults), preschool children's knowledge about animals' essential properties is less specific, perhaps consisting of only a *placeholder*—that is, a commitment to the existence of internal biological mechanisms without any notion of what those mechanisms might be (Gelman & Wellman, 1991; Johnson & Solomon, 1997; Medin & Ortony, 1989). But despite these differences, we suggest that an underlying feature that is inferred from observable ones can affect classification regardless of whether it is defining, truly essential, or only strongly associated with the category.

This fact is important, because not all categories may be essentialized to the same degree or at all. For example, although underlying causal properties might be important for complex artifacts (e.g., automobiles, computers), simple artifacts like pencils and wastepaper baskets appear to be defined more in terms of their perceptual and/or functional properties (Chaigneau, Barsalou, & Sloman, 2004; Malt, 1994; Malt & Johnson, 1992, 1998; cf. Bloom, 1998; Matan & Carey, 2001; Rips, 1989). Even for biological kinds, people may believe that individual animals can vary in the degree to which they participate in their kinds' essential properties and processes (Gelman & Hirschfeld, 1999). Consistent with this interpretation, Hampton (1995) demonstrated that even when a biological category's so-called essential properties are unambiguously present (or absent) in an individual, its characteristic features continue to exert an influence on judgments of category membership (also see Braisby, Franks, & Hampton, 1996; Kalish, 1995). But because the diagnostic reasoning view only requires that underlying features be strongly associated with, rather than essential to, category membership, it applies to even these cases in which fully essential features may be absent.

The diagnostic reasoning view provides a framework in which to understand the mixed results obtained across transformation studies. As was previously mentioned, well-known studies such as those by Keil (1989) and Rips (1989) have shown that animals that undergo transformations so that they look like another species are usually judged to have not undergone a change in category memberships. On one hand, these results suggest that, although observable features can serve as evidence for underlying ones under normal conditions, people know that such causal inferences may be unjustified when those features are transformed through external intervention (Strevens, 2007). Even though gender cues such as body shape and voice pitch are biologically determined, they can be manipulated (transformed) by external interventions such as surgery and the ingestion of hormones; even though raccoons normally have gray, mottled fur, these properties can be modified by nefarious veterinarians. In such circumstances, people may simply rely on the underlying (and perhaps essential) properties whose presence was inferred *before* the transformation to determine that category membership remains unchanged, despite new appearances.

But more recent evidence suggests that people reason diagnostically to category membership even for cases involving transformations. In a replication of Rips (1989), Hampton et al. (2007) found that whether a transformed animal was judged to have changed category membership often depended on what the participants could infer about underlying causal processes and structures. The participants in these studies were told a story about, for example, a bird that had normal bird-like features (it had feathered wings, ate seeds, lived in a nest in a tree, etc.) until it was exposed to hazardous chemicals, after which it mutated to take on insect-like properties (a brittle outer shell, transparent membranes for wings, etc.). As in Rips's study, a (small) majority of participants in Hampton et al. judged the transformed animal to still be a bird, whereas a (large) minority judged that it was now an insect. But although the judgments of the latter group (dubbed the *phenomenalists* by Hampton et al.) would seem to be based on the animals' appearance, the justifications that they provided for their choices indicated instead that many used the animals' new properties to infer deeper changes. For example, the participants assumed that a giraffe that lost its long neck also exhibited new behaviors that were driven by internal changes (e.g., to its nervous system), which in turn signaled a change in category membership (to a camel). Conversely, those participants who judged that the transformed animal's category was unchanged (the *essentialists*) often appealed to the fact that it produced offspring from its original category, from which they inferred the *absence* of important internal changes (e.g., the animal's DNA was unchanged). In other words, rather than the (so-called) phenomenalists' using only observable features, and rather than the essentialists' relying just on the presence of previously inferred underlying properties, both groups used observable features to infer the state of internal causal structures and processes and decided category membership on that basis.

Other studies provide indirect evidence for the presence of diagnostic reasoning during classification. For example, recall that the causal status effect is the phenomenon in which features earlier in a causal chain are weighed more heavily than later features. On one hand, Ahn and her colleagues have frequently attributed this effect to the fact that the early features have more dependents (i.e., effects, features that depend on them; Ahn & Kim, 2001; Ahn et al., 2000; Kim & Ahn, 2002a, 2002b; Sloman et al., 1998). However, the causal status effect has also been attributed to essentialism (Ahn & Kim, 2001; Ahn et al., 2000; Rehder, 2003b). For example, Rehder (2003b) proposed that, when category features X, Y, and Z are related in a causal chain and the category is assumed to be essentialized, classifiers will assume that the category's causal model consists of the following chain: essence → X → Y → Z. Under these circumstances, the participants were likely to have reasoned backward from observable features to the essential properties, and, of course, features closer (in a causal sense) to the essence (e.g., X) were taken to be more diagnostic of that disease than more remote features (e.g., Z). Consistent with this interpretation, Rehder and Kim (2009) found a stronger causal status effect when observable features were given an explicit underlying cause (also see Rehder, 2003b). And, Ahn and her colleagues found that expert clinicians both view mental disorders as less essentialized than laypersons (Ahn, Flanagan, Marsh, & Sanislow, 2006) and exhibit only a weak causal status effect (Ahn, Levin, & Marsh, 2005), consistent with the idea that the causal status effect depends on diagnostic reasoning from observable features to underlying properties that are strongly associated with category membership.

Finally, in this article, we have emphasized the diagnostic reasoning from observed to underlying features. However, the causal reasoning that often underlies categorization can also be *prospective*, from the causes of an underlying feature, as when a physician diagnoses a disease on the basis of its potential causes. Evidence that people will reason causally either prospectively or diagnostically in an appropriate manner when deciding category membership is provided by studies in which the direction of the causal links relating observable features to underlying ones has been directly manipulated. For example, Rehder (2007) showed that an object's degree of category membership increased nonlinearly with its number of observable features when those features were effects (a result related to Experiment 5's discounting effect) as compared with the linear increase that obtained when those features were causes, results consistent with a normative account of causal reasoning (see also Chaigneau et al., 2004). As was previously mentioned, our lab has shown stronger inferences to an underlying feature in the forward (cause-to-effect) direction than in the backward (effect-to-cause) direction (Tversky & Kahneman, 1980). The distinction between diagnostic and prospective classification also affects how categories are learned. For example, Waldmann and his colleagues found that the standard blocking effect, in which initially learned cues inhibit the learning of subsequent cues, arises only when category features are construed as causes rather than effects (Waldmann, 2000; Waldmann & Holyoak, 1992) and that learning is facilitated to the extent that observable features exhibit the pattern of correlations that one would expect given the direction of the causal arrow (Waldmann et al., 1995). Of course, this evidence for diagnostic versus prospective classification adds to the already large body of evidence documenting the causal reasoning processes that constitute many category-based judgments. For example, asymmetries characteristic of causal reasoning have been established in category-based induction, both when features are inferred in individual category members (Rehder & Burnett, 2005) and when they are projected to a whole class of objects (Medin, Coley, Storms, & Hayes, 2003; Rehder, 2006, 2009; Rehder & Hastie, 2004).

In summary, an enduring problem in the field of categorization has been to account for the undisputed facts that everyday categorization is based on observable properties and that many categories appear to have an underlying reality that establishes category membership. Although previous demonstrations of the importance of underlying properties and structures to classification have involved objects that have undergone hypothetical transformations, they provide no reason to think that those beliefs play any role in acts of classifications involving objects displaying their normal features. To our knowledge, the present study is the first to directly demonstrate how beliefs about underlying properties can influence the classification of untransformed objects.

Implications for Computational Models

The evidence that categorization can involve probabilistic causal inference has several implications for models of theory-based effects in the psychology of concepts. In this section, we consider several theoretical models as accounts of these findings.

The dependency model. One model relevant to the present findings is the dependency model proposed by Sloman et al. (1998). The dependency model characterizes the theoretical knowledge that classifiers have about categories in terms of a network of dependency relations among category features, where a causal relation is one type of dependency relation (an effect depends on its causes). Given a category's network, the dependency model predicts that features will be weighed more heavily to the extent that they have more dependents (i.e., effects). This includes the features that they cause directly as well as those that they cause indirectly through other features.

Unfortunately, the dependency model's reliance on a feature's dependents makes it unable to account for the present results showing that feature importance varied with the presence of an extra cause. For example, given the category structures in Experiment 1 (Figure 1), the dependency model predicts that features A_1 and B_1 should be equally diagnostic of their respective categories (because they have an equal number of dependents—viz., zero), but, as we have seen, A_1's additional cause made it more diagnostic than B_1. This incorrect prediction adds to the dependency model's generally mixed record of empirical support. On one hand, testing natural categories including both artifacts (e.g., chairs, guitars) and biological kinds (e.g., apples, robins), Sloman et al. (1998) found positive correlations between features' number of dependents and their importance to category membership. On the other hand, Rehder and Kim (2006) systematically manipulated features' number of dependents and found no evidence that more dependents led to greater categorization weight. And, Rehder and Kim (2009) found a stronger causal status effect (1) when categories were essentialized and (2) when interfeature causal links were probabilistic versus deterministic, results that are both at odds with the predictions of the dependency model. Thus, we must look beyond the dependency model for a comprehensive account of the effect of causal knowledge on categorization.

The KRES model. Rehder and Murphy (2003) proposed the KRES recurrent connectionist model, which incorporates knowledge in the form of preexisting excitatory links between features. As a learning model, KRES accounts for a number of known effects of prior knowledge on category learning (Rehder & Murphy, 2003; Harris & Rehder, 2006). In addition, because excitatory links allow features to activate one another, KRES can model how observed features might activate unobservable ones that in turn activate the category label. For example, Rehder and Murphy demonstrated how KRES accounts for Murphy and Allopenna's (1994) finding that participants were able to correctly classify features that they had rarely seen before on the basis of relations between them and frequently observed features. It does so because the rare features activate the frequent ones that in turn activate the correct category label (see also Heit & Bott, 2000).

Unfortunately, however, because KRES only represents knowledge in the form of symmetrical excitatory links, it is unable to account for the numerous causal asymmetries that we have described, including the discounting effect observed in Experiment 5. For example, for that experiment, KRES would predict that A_1 should activate U_A as strongly as B_1 activates U_B, and thus classification of test item A_1B_1 should be at chance.

Causal-model theory and the generative model. Another approach to addressing the effect of interfeature causal relations in categorization is the general framework known as *causal-model theory* (Sloman, 2005; Waldmann, Hagmeyer, & Blaisdell, 2006; Waldmann & Holyoak, 1992). A causal model—a system of causally interrelated variables—exhibits the properties of causal graphical models, which specify how one should learn, reason with, and act on those variables (Glymour, 1998; Jordan, 1999; Pearl, 1988, 2000). Applying the causal model approach to classification, Rehder (2003a, 2003b; Rehder & Kim, 2006) proposed a generative model of classification in which features of a category are treated as variables in a causal model. Although causal graphical models themselves make no assumptions regarding the details of the causal relationships, the generative model assumes that interfeature causal links are represented as probabilistic causal mechanisms and that classifiers consider whether an object is likely to have been produced or generated by those causal mechanisms. Objects likely to have been generated by a category's causal model are considered to be good category members, and those unlikely to be generated are poor category members. An object is classified into the category that is most likely to have generated it (taking into account the categories' base rates).

There are two ways that the generative model can account for the result that observed features caused by underlying properties are more diagnostic of category membership. The first approach corresponds to the explicit causal reasoning account that we have described in this article. As a type of a causal graphical model, a category's network of interfeature causal links supports the elementary causal inferences required to account for the results in Experiments 1–5. Indeed, Rehder and Burnett (2005) confirmed that people are more likely to infer the presence of a cause feature when its effect was present (and vice versa). And, as was previously mentioned, they also exhibited the kind of discounting observed in the present Experiment 5: The presence of an effect's cause was rated as less certain when an alternative cause was present. Although Rehder and Burnett also observed some discrepancies from normative reasoning, current evidence indicates that people can readily engage in the causal reasoning from observed to unobserved features suggested by these experiments (also see Sloman & Lagnado, 2005; Waldmann & Hagmayer, 2005).

A second way that the generative model can account for the present results is that observed features caused by underlying properties are likely to be perceived as more prevalent among category members than other features, because they are more likely to be generated by the category's causal model. Suppose that category feature A_i is caused by feature A_j—specifically, that A_j produces A_i via some causal mechanism that operates with probability m_{ji} when A_j is present (and has no effect on A_i when A_j is absent). Under these assumptions, Rehder (2003b) derived A_i's probability among Category A members, $P(\mathrm{A}_i | \mathrm{C}_\mathrm{A})$:

$$P(\mathrm{A}_i | \mathrm{C}_\mathrm{A}) = m_{ji}P(\mathrm{A}_j | \mathrm{C}_\mathrm{A}) + b_i - m_{ji}b_iP(\mathrm{A}_j | \mathrm{C}_\mathrm{A}), \quad (1)$$

where b_i is the probability that A_i is brought about by alternative (background) causes associated with the category. In other words, the probability of feature A_i in some category member is the probability that it is either brought about by the causal mechanism [which is the probability that A_j is present times the probability that the causal mechanism operates, $m_{ji}P(\mathrm{A}_j | \mathrm{C}_\mathrm{A})$] or brought about by the background causes b_i.[4] When the cause A_j is a defining feature, $P(\mathrm{A}_j | \mathrm{C}_\mathrm{A}) = 1$ and Equation 1 reduces to

$$P(\mathrm{A}_i | \mathrm{C}_\mathrm{A}) = m_{ji} + b_i - m_{ji}b_i. \quad (2)$$

Equation 2 illustrates how the presence of a causal relationship between an underlying and an observable feature can make the underlying feature more prevalent in category members. For example, when $b_i = .75$ and there is no causal link between A_i and A_j ($m_{ji} = 0$), $P(\mathrm{A}_i | \mathrm{C}_\mathrm{A}) = .75$. But when, say, $m_{ji} = .75$, $P(\mathrm{A}_i | \mathrm{C}_\mathrm{A})$ increases to .94. The introduction of additional causes would make A_i even more probable [e.g., if A_i was independently caused by another defining feature, A_k, via a causal mechanism with $m_{ki} = .75$, $P(\mathrm{A}_i | \mathrm{C}_\mathrm{A})$ would increase to .98]. A simple example follows: You may have a rough estimate of the number of homeless people in the U.S. and an intuition about why homelessness occurs (e.g., lack of ambition), but if you were to learn that there are many additional causes of homelessness (e.g., poor education, mental health problems), you are likely to realize that homelessness is more prevalent than you previously thought. Similarly, if you learn that an observed category feature is causally related to an underlying property, you are likely to assume that the feature is more prevalent among category members. In terms used in the categorization literature, the feature's *category validity* (the probability of the feature given the category) has increased, and it is well known that higher category validity results in a feature's being more diagnostic (Rosch & Mervis, 1975). In the Appendix, we demonstrate how the generative view and its assumption regarding increased category validity for causally related features provides an account of each of the our experiments.

Of course, there are conditions under which a known cause might not lead to an increase in category validity. For example, you might have high confidence in your current estimate of homelessness, in which case, learning about additional causes might either lower the strength of the causes that you already knew about or be interpreted as the background causes about which you were previously ignorant (and, thus, the introduction of new causes might be accompanied for by a reduction in b_i in Equation 1). However, previous research provides good reason to expect that an explicit cause is likely to increase the subjective frequency of an event. For example, according to Tversky and Koehler's (1994) support theory, the subjective probability of an event increases when supporting evidence is enumerated (death due to cancer, heart disease, or some other natural cause) rather than summarized (death due to natural cause; see also Fischoff, Slovic, & Lichtenstein, 1978). And, Rehder and Milovanovic (2007) found that an event was rated as more probable as its number of causes increased (from 1 to 2 to 3). Consistent with this greater subjective probability, Rehder and Kim (2006) found that features with three causes were more diagnostic of their category than those with only one cause (also see Rehder, 2003a, and Rehder & Hastie, 2001). Finally, Rehder and Kim (2009) directly demonstrated that experimental manipulations that affected features' diagnosticity were accompanied by changes in their subjective category validity, further suggesting that changes in the former are mediated by the latter.

Additional research will be required to determine whether the greater diagnosticity of causally generated features arises from the explicit causal inferences that they support or their greater perceived category validity (or both). Note that these two accounts correspond to alternative views of what features constitute the models of our experimental categories. On the causal reasoning account, the causal models consist solely of the underlying features U_A and U_B, and participants reason explicitly from variables that are external to those models (the observable features) to the underlying ones. On the second account, the model includes both underlying and observable features (and the model generates the observable features with greater probability when they are linked to the underlying ones).

Although these two applications of the generative model are each sufficient to account for the present results, they are associated with distinct mental processes, and there are experimental designs that can distinguish between them.

Summary

We have demonstrated how features become more diagnostic of category membership to the extent that they are viewed as causally generated by underlying properties or structures. We characterized these effects as acts of causal reasoning in which classifiers reason causally from observable to unobserved internal features. We also demonstrated how the results can be accounted for by assuming that causally generated features are viewed as being more prevalent among category members than are other features.

AUTHOR NOTE

We thank Harlan Harris, Aaron Hoffman, Gregory L. Murphy, and three anonymous reviewers for their comments on an earlier draft of this article. Correspondence concerning this article should be addressed to B. Rehder, Department of Psychology, 6 Washington Place, New York, NY 10003 (e-mail: bob.rehder@nyu.edu).

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., Flanagan, E., Marsh, J. K., & Sanislow, C. (2006). Beliefs about essences and the reality of mental disorders. *Psychological Science*, **17**, 759-766.

Ahn, W., & Kim, N. S. (2001). The causal status effect in categorization: An overview. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 40, pp. 23-65). San Diego: 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, W., Levin, S., & Marsh, J. K. (2005). Determinants of feature centrality in clinicians' concepts of mental disorders. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Bloom, P. (1998). Theories of artifact categorization. *Cognition*, **66**, 87-93.

Braisby, N., Franks, B., & Hampton, J. (1996). Essentialism, word use, and concepts. *Cognition*, **59**, 247-274.

Chaigneau, S. E., Barsalou, L. W., & Sloman, S. A. (2004). Assessing the causal structure of function. *Journal of Experimental Psychology: General*, **133**, 601-625.

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review*, **104**, 367-405.

Diesendruck, G. (2001). Essentialism in Brazilian children's extensions of animal names. Developmental Psychology, **37**, 49-60.

Fischoff, B., Slovic, P., & Lichtenstein, S. (1978). Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception & Performance*, **4**, 330-344.

Gelman, S. A. (2003). *The essential child: The origins of essentialism in everyday thought*. New York: Oxford University Press.

Gelman, S. A., & Hirschfeld, L. A. (1999). How biological is essentialism? In D. L. Medin & S. Atran (Eds.), *Folk biology* (pp. 403-446). Cambridge, MA: MIT Press.

Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings of the nonobvious. *Cognition*, **38**, 213-244.

Glymour, C. (1998). Learning causes: Psychological explanations of causal explanation. *Minds & 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.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning & Verbal Behavior*, **18**, 441-461.

Hampton, J. A. (1995). Testing the prototype theory of concepts. *Journal of Memory & Language*, **34**, 686-708.

Hampton, J. A., Estes, Z., & Simmons, S. (2007). Metamorphosis: Essence, appearance, and behavior in the categorization of natural kinds. *Memory & Cognition*, **35**, 1785-1800.

Harris, H. D., & Rehder, B. (2006). Modeling category learning with exemplars and prior knowledge. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1440-1445). Mahwah, NJ: Erlbaum.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review*, **7**, 569-592.

Heit, E., & Bott, L. (2000). Knowledge selection in category learning. In D. L. Medin (Ed.), *The psychology of learning and motivation*, **39**, (Vol. 39, pp. 163-199). San Diego: Academic Press.

Hirschfeld, L. A. (1996). *Race in the making: Cognition, culture, and the child's construction of human kinds*. London: MIT Press.

Johnson, S. C., & Solomon, G. E. A. (1997). Why dogs have puppies and cats have kittens: The role of birth in young children's understanding of biological origins. *Child Development*, **68**, 404-419.

Jordan, M. I. (Ed.) (1999). *Learning in graphical models*. Cambridge, MA: MIT Press.

Kalish, C. W. (1995). Essentialism and graded category membership in animal and artifact categories. *Memory & Cognition*, **23**, 335-353.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Kelley, H. H. (1973). The process of causal attribution. *American Psychologist*, **28**, 107-128.

Kim, N. S., & Ahn, W. (2002a). Clinical psychologists' theory-based representations of mental disorders affect their diagnostic reasoning and memory. *Journal of Experimental Psychology: General*, **131**, 451-476.

Kim, N. S., & Ahn, W. (2002b). The influence of naive causal theories on lay concepts of mental illness. *American Journal of Psychology*, **115**, 33-65.

Malt, B. C. (1994). Water is not H_2O. *Cognitive Psychology*, **27**, 41-70.

Malt, B. C., & Johnson, E. C. (1992). Do artifacts have cores? *Journal of Memory & Language*, **31**, 195-217.

Malt, B. C., & Johnson, E. C. (1998). Artifact category membership and the intentional-historical theory. *Cognition*, **66**, 79-85.

Marsh, J., & Ahn, W. (2006). The role of causal status versus inter-feature links in feature weighting. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 561-566). Mahwah, NJ: Erlbaum.

Matan, A., & Carey, S. (2001). Developmental changes within the core of artifact concepts. *Cognition*, **78**, 1-26.

McCloskey, M., & Glucksberg, S. (1978). Natural categories: Well-defined or fuzzy sets? *Memory & Cognition*, **6**, 462-472.

McClure, J. (1998). Discounting causes of behavior: Are two reasons better than one? *Journal of Personality & Social Psychology*, **74**, 7-20.

McNorgan, C., Kotack, R. A., Meehan, D. C., & McRae, K. (2007). Feature–feature causal relations and statistical co-occurrences in object concepts. *Memory & Cognition*, **35**, 418-431.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review*, **10**, 517-532.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179-196). Cambridge: Cambridge University Press.

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

Morris, M. W., & Larrick, R. P. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review*, **102**, 331-355.

Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 904-919.

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: Cambridge University Press.

Rehder, B. (2003a). Categorization as causal reasoning. *Cognitive Science*, **27**, 709-748.

Rehder, B. (2003b). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **29**, 1141-1159.

Rehder, B. (2006). When causality and similarity compete in category-based property induction. *Memory & Cognition*, **34**, 3-16.

Rehder, B. (2007). Essentialism as a generative theory of classification. In A. Gopnik & L. Schultz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 190-207). Oxford: Oxford University Press.

Rehder, B. (2009). Causal-based property generalization. *Cognitive Science*, **33**, 301-343.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, **50**, 264-314.

Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General*, **130**, 323-360.

Rehder, B., & Hastie, R. (2004). Category coherence and category-based property induction. *Cognition*, **91**, 113-153.

Rehder, B., & Kim, S. (2006). How causal knowledge affects classification: A generative theory of categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **32**, 659-683.

Rehder, B., & Kim, S. (2009). *Causal status and coherence in causal-based categorization*. Manuscript submitted for publication.

Rehder, B., & Milovanovic, G. (2007). Bias toward sufficiency and completeness in causal explanations. In D. MacNamara & G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (p. 1843). Mahwah, NJ: Erlbaum.

Rehder, B., & Murphy, G. L. (2003). A Knowledge-Resonance (KRES) model of category learning. *Psychonomic Bulletin & Review*, **10**, 759-784.

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.

Rips, L. J. (2001). Necessity and natural categories. *Psychological Bulletin*, **127**, 827-852.

Rosch, E. H., & Mervis, C. B. (1975). Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology*, **7**, 573-605.

Ross, B. H. (1996). Category representations and the effects of interacting with instances. *Journal of Experiment Psychology: Learning, Memory, & Cognition*, **22**, 1249-1265.

Ross, B. H. (1997). The use of categories affects classification. *Journal of Memory & Language*, **37**, 240-267.

Ross, B. H. (1999). Postclassification category use: The effects of learning to use categories after learning to classify. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **25**, 743-757.

Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives*. Oxford: Oxford University Press.

Sloman, S. A., & Lagnado, D. A. (2005). Do we "do"? *Cognitive Science*, **29**, 5-39.

Sloman, S. A., Love, B. C., & Ahn, W. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, **22**, 189-228.

Strevens, M. (2007). Why represent causal relations? In A. Gopnik & L. Schultz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 245-260). Oxford: Oxford University Press.

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

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, **101**, 547-567.

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

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 216-227.

Waldmann, M. [R.], Hagmayer, Y., & Blaisdell, A. (2006). Beyond the information given: Causal models in learning and reasoning. *Psychological Science*, **15**, 307-311.

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.

NOTES

1. Interestingly, the negative items were classified into the predicted category significantly less frequently than the positive items (.84 and .68, respectively). However, this result was not replicated in subsequent experiments and so will not be discussed further.

2. We thank Michael Strevens for suggesting the design of this experiment.

3. Note that a potential concern is that in Experiment 5 participants were reasoning not on the basis of the causal relations per se but rather with a superficial encoding of the statistical information that accompanied the background causes. To continue the example from above, for the feature "hyperactive immune system," participants may have only encoded "a hyperactive immune system occurs in 50% of ants that don't have blood high in iron sulfate"; for "thick blood" they may have only encoded "ants that don't have blood high in iron sulfate never have thick blood." That is, the categories might have been encoded without any causal information at all. Combined with the explicit feature base rates (that features occur in "most" category members), this information alone might account for participants' choices.

We present two responses to this concern. The first is to note that this alternative account is consistent with performance on the positive but not on the negative items. For example, suppose that because features A_1, A_2, B_1, and B_2 were described as occurring in "most" of their respective category members, one assumes $P(A_1 | C_A) = P(A_2 | C_A) = P(B_1 | C_B) = P(B_2 | C_B) = .75$. Also, suppose that on the basis of statistical information that accompanied the description of background causes, one assumes that $P(A_1 | {\sim}C_A) = P(B_2 | {\sim}C_B) = .50$ and $P(A_2 | {\sim}C_A) = P(B_1 | {\sim}C_B) = 0$. Given these quantities, the probability that test item A_1B_1 is a member of Category A rather than of Category B, $P(C_A | A_1B_1)$, is equal to $P(A_1B_1 | C_A)/[P(A_1B_1 | C_A) + P(A_1B_1 | C_B)] = P(A_1 | C_A)P(B_1 | C_A)/[P(A_1 | C_A)P(B_1 | C_A) + P(A_1 | C_B)P(B_1 | C_B)] = P(A_1 | C_A)P(B_1 | {\sim}C_B)/[P(A_1 | C_A)P(B_1 | {\sim}C_B) + P(A_1 | {\sim}C_A)P(B_1 | C_B)] = (.75)(0)/[(.75)(0) + (.50)(.75)] = 0$. Using similar reasoning, $P(C_A | A_2B_2) = (.75)(.50)/[(.75)(.50) + (0)(.75)] = 1$. That is, the positive test items are classified as members of Category B and of Category A, respectively, consistent with the results in Table 6. In contrast, however, whereas for the negative test items $P(C_A | {\sim}A_1{\sim}B_1) = (.25)(1)/[(.25)(1) + (.50)(.25)] = .67$ and $P(C_A | {\sim}A_2{\sim}B_2) = (.25)(.50)/[(.25)(.50) + (1)(.25)] = .33$, Table 6 indicates that the participants were at chance on those items.

The second response is that, to definitively rule out the possibility that the results in Experiment 5 were due to use of statistical information rather than to causal reasoning, we ran another version of Experiment 5 in which statistical information was omitted. For example, the participants were simply told that "besides blood high in iron sulfate, there are also one or more unknown causes of hyperactive immune system" and "besides blood high in iron sulfate, there are no other causes of thick blood." The results were qualitatively identical to those in Table 6, again suggesting that the participants were not solely reasoning with the probabilistic information that accompanied the causal relations. We thank an anonymous reviewer for raising this issue.

4. Equation 1 characterizes what is known in the Bayesian network literature as a *fuzzy-or* network, in which two causes (in this case, A_j and the background cause) are independent and can each bring about the effect (A_i). This formalization is equivalent to the one presented in Cheng's Power PC theory of causal induction (Cheng, 1997; Glymour & Cheng, 1998).

APPENDIX

Table A1 presents an example of the generative model's quantitative predictions for Experiments 1–5, assuming that the effect of causal links is to make features more prevalent in category members. For each experiment, the table presents the probability of each feature and each test item in each category and the probability of classifying a test item into Category A. To compute each feature's probability within a category (i.e., its category validity), the following assumptions are made (unless otherwise stated in a specific experiment): (1) If not involved in a causal relationship, a feature appears in category members with probability .75 and in members of the contrast category with probability .10. (2) When causal links are present, they operate with probability .75 ($m = .75$), and the probability of the background cause operating is .25 ($b = .25$), and the effect feature's category validity is given by Equation 1. Note that although these assumptions are necessary to provide a quantative example, the generative model's qualitative predictions regarding each test item's category membership hold for all parameter values so long as a feature has a higher category validty when it is caused by an underlying feature (and that certain boundary conditions involving probabilities of 0 and 1 are avoided; for example, the choice probabilities for some test items become undefined if the probability of a feature in the other category is defined as 0 or 1).

Table A1
Predictions of the Generative Model for Experiments 1–5

	Experiment									
	1		2		3		4		5	
	Category A	Category B	Category A	Category B	Category A	Category B	Category A	Category B	Category A	Category B
	Probability of the Feature Given the Category (Category Validity)									
A_1	.813	.100	.813	.100	.925	.100	.813	.100	.875	.500
A_2	.750	.100	.707	.100	.700	.100	.813	.100	.750	0
B_1	.100	.750	.100	.750	.100	.700	.100	.672	0	.750
B_2	.100	.813	.100	.673	.100	.925	.100	.672	.500	.875
	Probability of the Test Item Given the Category									
A_1B_1	.081	.075	.081	.075	.093	.070	.081	.067	0	.375
A_2B_2	.075	.081	.071	.067	.070	.093	.081	.067	.375	0
$\sim A_1 \sim B_1$	.169	.225	.169	.225	.068	.270	.169	.295	.125	.125
$\sim A_2 \sim B_2$	.225	.169	.264	.295	.270	.068	.169	.295	.125	.125
	Probability That Test Item Is a Member of Category A									
A_1B_1	.520 (A)		.520 (A)		.569 (A)		.547 (A)		0 (B)	
A_2B_2	.480 (B)		.513 (A)		.431 (B)		.547 (A)		1 (A)	
$\sim A_1 \sim B_1$	.429 (B)		.429 (B)		.200 (B)		.364 (B)		.500	
$\sim A_2 \sim B_2$	.571 (A)		.472 (B)		.800 (A)		.364 (B)		.500	

Note—The participants' category choices are presented in parentheses.

The probability of a test item A_iB_j within Category C_k is computed by multiplying the category validities of the individual features—that is, $P(A_iB_j \mid C_k) = P(A_i \mid C_k)P(B_j \mid C_k)$. Finally, the probability of a test item's being a member of Category A, $P(C_A \mid A_iB_j)$, is computed from $P(A_iB_j \mid C_A)$ and $P(A_iB_j \mid C_B)$, according to Bayes's law,

$$P(C_A \mid A_iB_j) = P(A_iB_j \mid C_A)P(C_A)/[P(A_iB_j \mid C_A)P(C_A) + P(A_iB_j \mid C_B)P(C_B)]. \qquad \text{(A1)}$$

We assume that the prior probabilities of the categories are equal—that is, $P(C_A) = P(C_B) = .5$.

Table A1 indicates that the generative model reproduces the qualitative results of all five experiments. For Experiment 1, it predicts that causally generated features A_1 and B_2 have higher category validity than the unrelated features A_2 and B_1 (.813 vs. .750). Because test item A_1B_1 thus provides stronger evidence for Category A than for Category B, it is classified as a member of Category A; because A_2B_2 provides stronger evidence for Category B than for Category A, it is classified as a member of Category B. In Experiment 2, because of its indirect to link to U_A, A_2 is generated more reliably by Category A than B_2 is by Category B (.707 vs. .673); thus, the crucial test item, A_2B_2, is classified as a member of Category A. The predictions for Experiment 3 were generated assuming that $m = .90$ for A_1 and B_2 and $m = .60$ for A_2 and B_1, with the result that A_1 and B_2 have higher category validity than A_2 and B_1 (.925 vs. .700) (and, thus, A_1B_1 is classified as a member of Category A and A_2B_2 is classified as a member of Category B). The predictions for Experiment 4 were generated from Equation 1, assuming that $P(U_A \mid C_A) = 1$ and $P(U_B \mid C_B) = .75$; the result is greater category validity for A_1 and A_2 than for B_1 and B_2 (.813 vs. .672), and, thus, both test items are classified as members of Category A. Finally, the predictions for Experiment 5 were generated from Equation 2, assuming $b = .5$ for A_1 and B_2 and $b = 0$ for A_2 and B_1. Moreover, the base rates of A_1 and B_2 in the opposing categories (B and A) are .5 and those of A_2 and B_1 are 0. The result is that test item A_1B_1 is a more probable member of Category B than of Category A (.375 vs. 0), whereas the reverse is true for A_2B_2. Consistent with the results of Experiment 5, no difference is predicted for the negative items.

(Manuscript received October 25, 2008;
revision accepted for publication February 18, 2009.)