COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13240

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13240

A Context-Dependent Bayesian Account for Causal-Based Categorization

Nicolás Marchant,[a] Tadeg Quillien,[b] Sergio E. Chaigneau[a,c]

[a]*Center for Social and Cognitive Neuroscience, School of Psychology, Universidad Adolfo Ibáñez*
[b]*School of Informatics, University of Edinburgh*
[c]*Center for Cognitive Research (CINCO), School of Psychology, Universidad Adolfo Ibáñez*

Received 25 June 2022; received in revised form 21 December 2022; accepted 30 December 2022

Abstract

The causal view of categories assumes that categories are represented by features and their causal relations. To study the effect of causal knowledge on categorization, researchers have used Bayesian causal models. Within that framework, categorization may be viewed as dependent on a likelihood computation (i.e., the likelihood of an exemplar with a certain combination of features, given the category's causal model) or as a posterior computation (i.e., the probability that the exemplar belongs to the category, given its features). Across three experiments, in combination with computational modeling, we offer evidence that categorization is better accounted for by assuming that people compute posteriors and not likelihoods, though both probabilities are closely related. This result contrasts with existing analyses of causal-based categorization, which assume that likelihood computations give a good approximation of human judgments. We also find that people are able to compute likelihoods in a closely related task that elicits judgments of consistency rather than category membership judgments. Our analyses show that people do use causal probabilistic information as prescribed by a Bayesian model but that they flexibly compute likelihoods or posteriors depending on the task. We discuss our results in relation to the relevant literature on the topic.

Keywords: Causality; Causal-based categorization; Bayesian reasoning; Computational modeling

1. Introduction

The view that categorization is causal-probabilistic reasoning has a long history (Ahn & Kim, 2001; Ahn, Kim, Lassaline, & Dennis, 2000; Gelman, 2003; Quillien, 2018; Rehder,

Correspondence should be sent to Nicolás Marchant, Avda. Presidente Errázuriz 3328, 7550313 Las Condes, Santiago, Chile. E-mail: nicolasmarchant@alumnos.uai.cl

2003a, 2003b, 2010, 2017; Waldmann & Hagmayer, 2006; Waldmann, Holyoak, & Fratianne, 1995). In this view, a category is represented by a causal model, with features playing one or more causal roles (e.g., having low-density bones *causes* birds to be light, which in turn *causes* them to be able to fly) and exemplars being represented by combinations of feature states (e.g., a bird that has denser bones and is less able to fly).

A normative account for probabilistic reasoning and human inferences comes from Bayesian causal models and graph theories (Gopnik & Wellman, 2012; Hagmayer, 2016; Holyoak & Cheng, 2011; Pearl, 2000; Rips, 2008; Rottman & Hastie, 2014; Sloman & Lagnado, 2015). Bayesian models have also been proposed as theories of categorization. Following Bayesian analyses of categorization (e.g., Anderson, 1991; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Rehder, 2017; Zhao, Lucas, & Bramley, 2021), it is possible to assume that the categorizer is trying to compute the probability that an object o belongs to category k, given that it has features $f = \{f1, f2\ldots\}$. That is, the categorizer uses the evidence revealed by the observation of the object features to compute a posterior probability $p(k|o_f)$. This posterior probability can be computed using Bayes' rule:

$$P\left(k|o_f\right) = \frac{p(o_f|k)P\,(k)}{P\left(o_f\right)}, \tag{1}$$

where $p(o_f\,|k)$ is the likelihood, that is, the probability of observing that the object has features f given that the object belongs to category k; $p(k)$ is the prior probability of observing a category k; and $p(o_f)$ is the prior probability that one would observe the object to have those features across all possible categories (i.e., k and $\neg k$).

While a Bayesian analysis suggests that people compute the posterior probability $p(k|o_f)$, a large amount of work on causal-based categorization assumes that people's category membership judgments track the likelihood $p(o_f\,|k)$, with the proviso that in the typical experimental condition, subjects are provided with only one possible category (the generative model, GM; Rehder, 2003a, 2003b, 2015; Rehder & Kim, 2006). This procedure could be sensible because the two measures (i.e., likelihoods and posteriors) are closely related. Even if we assume that category membership judgments involve computing $p(k|o_f)$, it seems reasonable to assume that people's judgments will vary as a function of $p(o_f\,|k)$. Modeling a categorizer who computes likelihoods requires fewer assumptions (on the part of the modeler) than modeling a categorizer who computes posterior probabilities, so a likelihood-based model of categorization is an attractive one. However, this strategy is not necessarily perfect, and we find that in causal-based categorization tasks, likelihoods are not guaranteed to always perfectly track posterior probabilities (see the Supporting Information).

Therefore, empirical results showing that people's categorization judgments track the likelihood $p(o_f\,|k)$ do not necessarily show that causal-based categorization involves full Bayesian inference. That is, it remains possible that in a case where likelihoods and posteriors diverge, people's judgments track the posterior instead of the likelihood.

Here, we report a series of experiments that are designed to explore this issue and arbitrate between a likelihood-based and a posterior-based model of causal-based categorization. In three experiments, we present subjects with a simple A → B model and compare *category membership* judgments with a *consistency* condition, which entails judging the probability

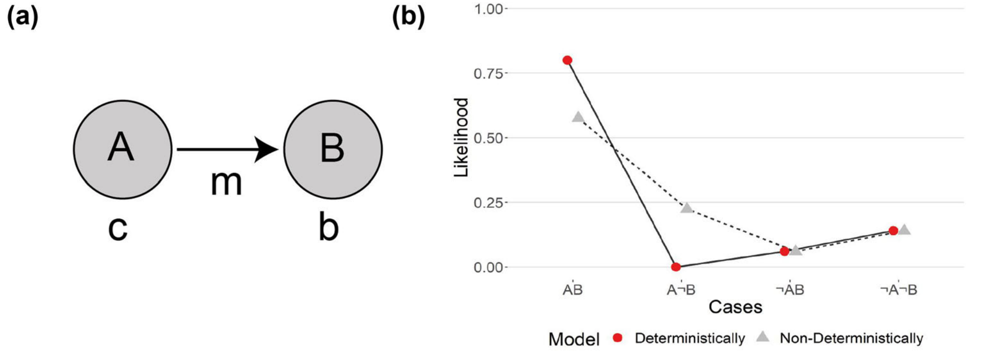

Fig. 1. Descriptive generative model (GM) predictions for a simple causal model.
Note: In panel (a), c = event A's base rate, b = event B's base rate, m = causal parameter. Panel (b) shows GM-predicted category likelihoods for a deterministically sufficient causal relation ($m = 1$) and for a non-deterministic relation ($m = 0.6$). Both model predictions are computed using base-rates values of $c = 0.8$ and $b = 0.3$, computed for all possible combinations of present and absent features (i.e., AB = both present; A¬B = A present and B absent; ¬AB = A not present and B present; ¬A¬B = neither A nor B are present).

of observing a certain pattern of features given category membership. Participants in the *consistency* question are essentially being asked about the likelihood $p(o_f|k)$. According to the GM, their pattern of judgments is expected to be close to that of participants in the *category membership* condition. However, we find that both judgments, although they are related, diverge in systematic ways.

We then explore a possible explanation for this divergence: While participants in the consistency condition are computing the likelihood $p(o_f|k)$, participants in the category membership condition are computing the posterior probability $p(k|o_f)$. Computational modeling offers support for this explanation.

2. The GM

As discussed above, the GM is a computational approach that formalizes Bayesian ideas in causal-based categorization (Rehder, 2003a, 2003b, 2015; Rehder & Kim, 2006). According to the GM, if people have a causal model about a category (e.g., that feature A causes feature B), then their categorization of an exemplar (i.e., a possible feature combination) is a function of the exemplar's estimated likelihood given the category's causal model, with likely exemplars being good category members and not likely exemplars being bad category members (Rehder, 2003a).

The simple A → B model can be specified by three parameters: the probability that A will be observed (c), the probability that B will be observed when its cause A is not present (b), and the causal strength linking A and B (m). Fig. 1 and Table 1 show a simple worked-out example. For a deterministically sufficient model (i.e., with $m = 1$), likelihoods can be computed by a conditional probability model, which may be more familiar to some readers

Table 1
Generative model (GM) likelihood equations for the simple causal models in Fig. 1

Exemplar (E)	L (E; c, m, b)	L (E; 0.8, 1, 0.3)	L (E; 0.8, 0.6, 0.3)
AB	$(c)(m + b - mb)$	0.8	0.58
A¬B	$(c)[(1 - m)(1 - b)]$	0	0.22
¬AB	$(1 - c)(b)$	0.06	0.06
¬A¬B	$(1 - c)(1 - b)$	0.14	0.14

Note. Likelihood (L) for each exemplar given GM parameters c (cause base rate), m (causal strength), and b (effect base rate). These likelihood equations are reported in Rehder (2003b). The right columns show an example for a specific combination of parameter values.

and thus aid in understanding (i.e., p(A), the cause's base rate; p(B), the effect's base rate; and $p(\text{B}|\text{A}) = 1$, the conditional probability of the effect given its cause). For non-deterministic models, though, the GM cannot be reduced to a conditional probability model.

In all three experiments we report here, we provided subjects with simple causal scenarios including one cause (A) and one effect (B) and asked them to perform one of two types of ratings (between subjects). As we will more fully describe shortly, one of our dependent variables was a *category membership* rating and the other was a *consistency* rating. Because the GM assumes that people compute exemplar likelihoods to guide their categorization judgments (with the same proviso discussed above), in the category membership condition, we asked subjects to perform category membership ratings. In the consistency rating condition, to capture likelihood judgments, we asked subjects to rate to what extent the exemplar was *expected*, if what they had learned regarding A causing B were correct.

2.1. Data analysis approach

For analyzing data in our three experiments, we used an individualized regression equation method that allows quantifying the relative contributions of individual features and of their causal relation (Rehder & Hastie, 2001). Because this method allows to parametrize individual features' and feature relations' contributions to category classification, it has been used to measure the magnitude of the coherence effect in causal categorization (Marchant & Chaigneau, 2020; Rehder & Kim, 2006, 2010), to measure the contribution of causal features over effect features (Marsh & Ahn, 2006; Rehder & Kim, 2006), and to measure the contribution of causally related properties in artifact categorization (Puebla & Chaigneau, 2014). In our experiments, we use it to gauge differences in response patterns for our two dependent measures.

In the individualized regression equation method, participants provide category membership ratings for all possible property combinations (AB, A¬B, ¬AB, ¬A¬B; recall that A is our cause and B our effect, across all our experiments), allowing the computation of individualized regression equations. When present and absent properties are coded, respectively, as 1 and -1 (i.e., effect coding in regression), these values can be entered into individualized regression equations to predict a participant's categorization ratings. Importantly for us, the two-way interaction term can be computed by entering the product of the corresponding property values as predictors into the equations. In our very simple A $\rightarrow$ B scenarios, the interaction term for scenarios with both properties present is $1 \times 1 = 1$, for those showing one property present and one absent it is 1 x $-1 = -1$, and for those showing both properties absent it is -1 x $-1 = 1$. Because individualized regression equations yield coefficients for single properties (A and B) and for the interaction term (AB), these regression coefficients can then be used as individual data points reflecting, across participants, the contribution of each predictor variable to the ratings. The distribution of coefficients across participants can then be submitted to significance tests. Importantly, the use of coefficients affords at least three advantages: The first is that coefficient values are statistically independent of each other (guaranteed because of experimental design reflected in effect coding) and of whether features are present or absent in the causal scenarios. Second, whereas raw ratings are generally not normally distributed, thus complicating statistical analyses, individualized regression coefficients tend to be normally distributed. Third, and importantly for us, the size of the AB interaction coefficient shows the relevance attributed to the causal relation in subjects' ratings considering all combinations of present or absent features (see Rehder & Kim, 2010).

2.2. *The coherence effect*

When contrasting both conditions, our focus will be on the contribution of individual features (A and B) and on the contribution of causal information (the AB interaction). This interaction operationalizes what is known as the coherence effect. The coherence effect in categorization occurs when people expect that features that are somehow related will tend to appear together in category exemplars (Hampton, Storms, Simmons, & Heussen, 2009; Malt & Smith, 1984; Murphy & Wisniewski, 1989; Wisniewski, 1995). If those expectations are at work, the coherence effect predicts that exemplars that preserve the expected correlations will be judged to be more likely given the causal model and relatively better exemplars than those that do not. In the GM, causal-probabilistic coherence is evidenced by the kind of nonlinearity in ratings illustrated in Fig. 1, which can be quantified with the AB interaction term in the individualized regression coefficients (i.e., if subjects do not take causal relations into account, the AB interaction term would be close to zero and smaller than the individual A and B features' coefficients and ratings would only linearly covary with the number of present/absent features weighted by their coefficients).

3. Experiments' overview

As may be clear by now, in our three experiments, participants were presented with scenarios describing a simple causal model with one cause and one effect (A $\rightarrow$ B). After the study phase, they were presented with exemplars showing all possible combinations of present and absent features (i.e., AB, $\neg$AB, A$\neg$B, $\neg$A$\neg$B). In all the experiments, the graphical causal model remained in sight as participants performed their ratings, to avoid the task becoming a memory task.

In the *consistency* condition, participants were asked to judge whether different feature combinations (i.e., AB, ¬AB, A¬B, ¬A¬B) were *expected* given a learned category's causal model (ranging from *definitely not expected* to *expected to only some degree* and to *definitely expected*). In contrast, in the *category membership* condition, we asked subjects to judge whether the exemplar being rated *belonged* to the learned category, by using a rating scale where the lower end of the scale was labeled *definitely does not belong* and the higher end of the scale was labeled *definitely does belong* (similar to rating scales used in Rehder, 2003a, 2003b; Rehder & Kim, 2006, 2010).

3.1. Experiment 1

This experiment tests if there are processing differences between the *consistency* and the *category membership* condition. Finding differences across conditions might allow extending the GM, given that there is no obvious mechanism that allows the GM to make different predictions for both experimental conditions. To test if there are differences, we used the A, B, and AB coefficients and evaluated if there were any important differences in the contributions of individual features and of the causal relation.

3.1.1. Participants

Twenty-four participants (18 females), all Spanish speakers aged 21 to 34 (mean = 24.92, *SD* = 3.02), voluntarily agreed to participate in the online experiment. Participants were randomly assigned to one of both task conditions (*consistency*, *category membership*) and experimental materials (exemplar order and type of scenarios, see below) to the constraint that an equal number of participants were in each cell. All participants read and accepted the informed consent approved by the Adolfo Ibáñez university ethics committee.

3.1.2. Design

In Experiment 1, we set up a 2 (task condition: *consistency*, *category membership*) x 3 (regression coefficient: A, B, and AB interaction) mixed design, with the last being a within-subjects factor. Participants learned about a simple causal model (A $\rightarrow$ B) and then used a rating scale (from 1 to 7) to categorize all possible combinations of present and absent features.

Participants in the consistency condition had to rate if the presented exemplar was expected given the studied category's causal model. In the category membership condition, participants had to perform a category membership rating.

3.1.3. Materials and procedures

Experiment 1 was implemented in the SurveyMonkey environment. Each online booklet described only one of two possible causal models (i.e., a scenario describing a type of rock or a scenario describing a type of neuropsychological disorder, see Table 2) and one of two possible rating questions (i.e., *consistency* or *category membership*). To promote a correct use of the rating scale, participants always rated the AB exemplar first, thus anchoring their responses on the high end of the scale. The other three scenarios were presented in one of

Table 2
Scenario type (Hamilton Disorder, Metamorphic Rock) by conditions (rating question)

Conditions	"Hamilton Disorder"	"Metamorphic Rock"
Consistency	If, as you learned about the "Hamilton Disorder," A causes B. Would you say that this case would be expected for a case of "Hamilton Disorder"?	If, as you learned about the "Metamorphic Rock," A causes B. Would you say that this case would be expected for a case of "Metamorphic Rock"?
Category membership	If, as you learned about the "Hamilton Disorder," A causes B. Would you say that this case belongs to the "Hamilton Disorder" category?	If, as you learned about the "Metamorphic Rock," A causes B. Would you say that this case belongs to the "Metamorphic Rock" category?

Table 3
Description of features by category names

Feature	"Hamilton Disorder"	"Metamorphic Rock"
A	FOX1 gene mutation	High calcium concentration
B	Difficulties to develop normal language	Being soft

three different Latin-square orders (order 1: ¬A¬B, A¬B, and ¬AB; order 2: ¬AB, ¬A¬B, and A¬B; order 3: A¬B, ¬AB, and ¬A¬B).

Each online booklet had a total of six pages. The first page contained the informed consent that every participant read and signed. The second page contained a cover story that described a category and its causal model (i.e., *Hamilton Disorder* was a type of neuropsychological disorder and *Metamorphic Rock* was a type of a rock). Both scenarios described a simple causal model with one cause and one effect (see Table 3). The same second page presented the instructions regarding rating-scale use. In the third to sixth pages, online booklets presented cases with different feature combinations, making a total of four cases. Each case described a researcher who found and described the scenario, and participants were asked to report their judgments by using the seven-point rating scale. Scale labels differed for each condition at the endpoints and at the middle point (category membership ranged from: *definitely does not belong* to *belongs only to some degree* to *definitely does belong*; consistency ranged from: *definitely not expected* to *expected to only some degree* to *definitely expected*).

The scenarios described in Table 3 read as follows. The "Hamilton Disorder" scenario: "There is a type of neuropsychological disorder known as Hamilton Disorder. In some cases, this type of disorder shows a mutation in the FOX1 gene. In some cases, it shows difficulties in normal language development. Not everyone who suffers from Hamilton Disorder has a FOX1 gene mutation nor develops difficulties in normal language development. However, when the FOX1 gene mutation is present, it causes those who suffer from Hamilton Disease to have difficulties in developing normal language because the FOX1 gene's molecular

Table 4
Raw average ratings for each condition and each case

	Cases			
Condition	AB	A¬B	¬AB	¬A¬B
Consistency	6.08 (1.31)	2.0 (1.35)	2.67 (1.23)	4.83 (1.85)
Category membership	5.42 (1.38)	2.08 (1.31)	2.92 (1.88)	3.08 (1.24)

Note. Standard deviations (*SD*) are in parentheses.

composition impairs the proteins required for developing normal language.” The “Metamorphic Rock” scenario: “There is a type of a mountain rock known as Metamorphic Rock. In some cases, this type of rock presents a high concentration of calcium, and in some cases, it becomes very soft. Not all Metamorphic rocks have a high calcium concentration, nor do all of them become soft. However, when high calcium concentrations are present, they cause Metamorphic Rocks to become soft, because the high calcium concentration decreases rock solidification.”

3.1.4. Results

All statistical analyses were performed using SPSS 26 and RStudio. Open data materials can be found at the end of the document. As mentioned earlier, we used an individualized multiple regression method that allowed us to measure the effect of individual features (the A and B coefficients) and coherence effects’ magnitude (the AB interaction coefficient), considering all four AB, A¬B, ¬AB, and ¬A¬B ratings. Table 4 shows the raw average ratings for each condition and for each feature-states combination.

We submitted the individualized regression coefficients to a 2 (task condition: *consistency*, *category membership*) x 3 (coefficient: A, B, and AB) mixed ANOVA, with the last being the repeated measures factor. We collapsed both scenarios (i.e., "Hamilton Disorder", "Metamorphic Rock") as we found no statistical differences between them and the scenarios factor did not interact with other factors in the design. Visual inspection of coefficients’ distributions and Shapiro–Wilk tests showed no significant deviations from normality. The analysis showed no main effect of task condition ($F = 0.09$, $p = .77$), a main effect of coefficient ($F(2,44) = 10.89$, $MSe = 0.51$, $p < .001$, $\eta_p^2 = 0.33$, power = 0.99), and a significant two-way interaction ($F(2,44) = 3.60$, $MSe = 0.51$, $p = .04$, $\eta_p^2 = 0.14$, power = 0.64). To follow up on the significant interaction, we performed planned comparisons across the repeated-measures coefficients factor. For the consistency condition, we found a significant difference when contrasting the AB interaction with the average of A and B main effects (AB greater than the average of A and B; $F(1,11) = 20.12$, $MSe = 0.93$, $p = .001$, $\eta_p^2 = 0.65$, power = 0.98) and a marginally significant difference between the A and B main effects ($F = 4.0$, $p = .07$). For the category membership condition, we did not find a significant difference between AB interaction term and the average of A and B main effects ($F = 1.37$, $p = .27$) nor did we find a significant difference between main effects ($F = 1.36$, $p = .27$). Fig. 2a illustrates our findings. Fig. 2b,c illustrates the AB interaction effect in a different manner. As shown by

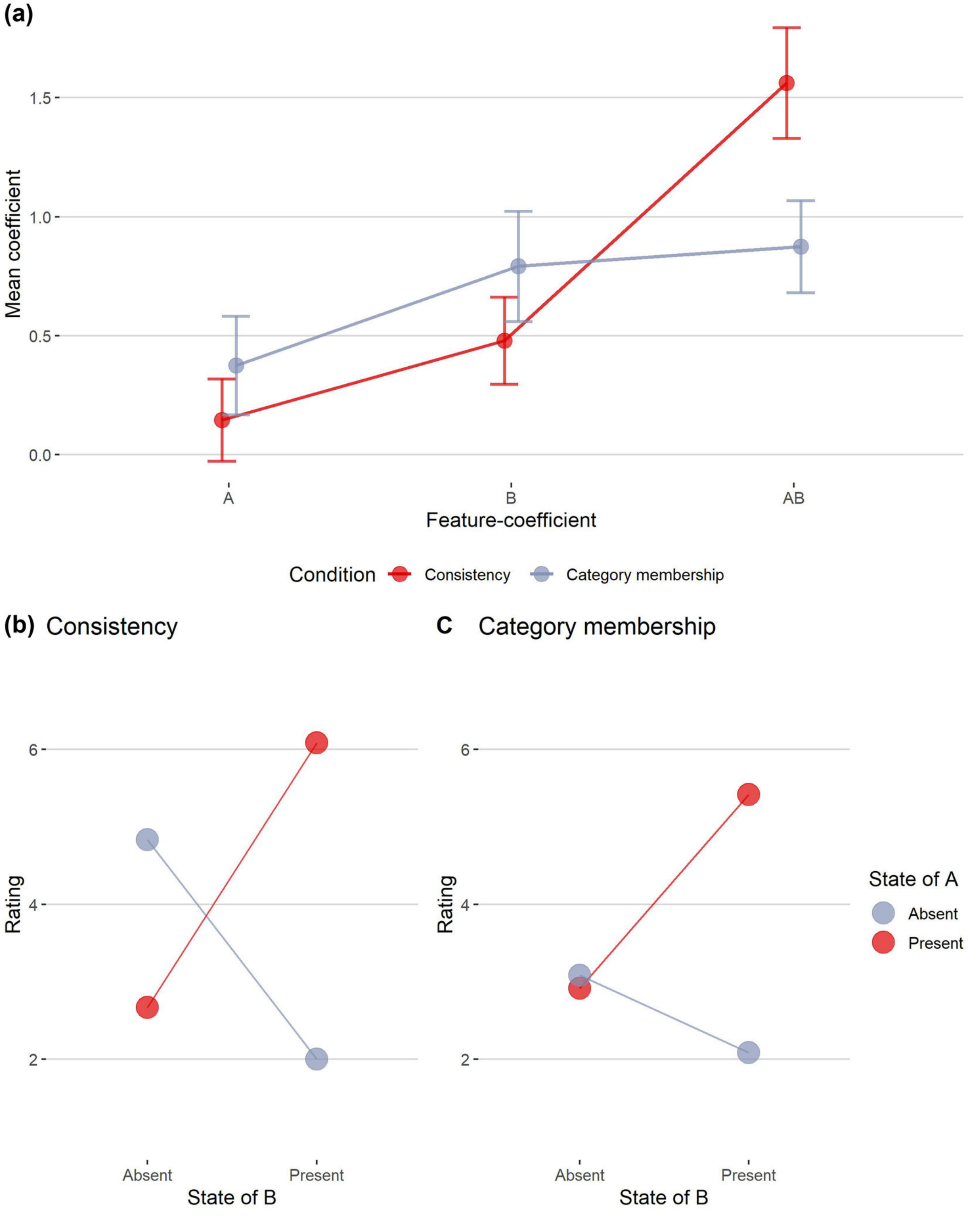

Fig. 2. In Experiment 1, mean regression coefficient plot and case feature interaction plots.
Note. Panel (a) shows the mean of regression coefficients of individualized regressions (A, B, and AB interaction) and by condition (*consistency* and *category membership*). Error bars are SE for the mean. Panel (b): crossover interaction effect of A and B causally related features in the consistency condition. Panel (c) crossover interaction effect of A and B causally related features in the category membership condition.

the crossover interaction, both conditions show a coherence effect, but the size of the effect is clearly larger for consistency than for category membership. This is evident in Fig. 2b where subjects in the consistency condition rated the ¬A¬B case with a higher category rating value than in the category membership condition (Fig. 2c).

3.1.5. *Discussion*

As our analysis predicted, subjects in the consistency condition responded in a different pattern than in the category membership condition. In the consistency condition, subjects were heavily influenced by the causal relation and less by the individual features, whereas in the category membership condition, they were influenced by the causal relation but only to approximately the same level as they were influenced by the individual features. The GM model does not naturally account for this divergence.

There are at least two possible accounts for these results. On one account, it is possible that subjects in the category membership condition were responding based on similarity. It has been argued that when people respond based on similarity, they tend to disregard probabilistic information (Barbey & Sloman, 2007; Sloman & Lagnado, 2015). Thus, they might have not considered the category membership rating question as a causal-probabilistic problem. For example, they could have conceptualized the individual features and the causal relation as three independent variables—with the causal relation becoming a configural cue—combined in an associative computation (e.g., Gluck & Bower, 1988; Gluck & Myers, 2001, Marchant, Canessa, & Chaigneau, 2022). This would account for the lower weight of causal information and higher relative weight of featural information in the category membership condition.

A different account that we wish to consider is the following. In the consistency condition, participants are computing the likelihood of observing the features associated with the observed exemplar, assuming that the exemplar belongs to the focal category. By contrast, in the category membership condition, participants are computing the posterior probability that the exemplar belongs to the focal category, given its observed features. As we explain in Section 4, this hypothesis predicts the (empirically observed) lower reliance on causal coherence in the category membership condition, compared to the consistency condition. Briefly, if participants assume that category-specific features are rare in non-category members (see Hampton, 2006; Jones, 1983; Rosch & Mervis, 1975), then we expect that individual features are more diagnostic for category membership than they are for consistency judgments.

3.2. *Experiment 2*

In this second experiment, we aimed at replicating our results in Experiment 1 regarding individual features being as important as the causal relation in the category membership condition but less important in the consistency condition. Also, we aimed at discarding the similarity account that holds that when categorizing, people use similarity-based processing, which would explain the differences between the category membership and consistency conditions. In Experiment 2, we provided explicit information about the strength of the causal link. If the similarity account is correct, then subjects should disregard causal-probabilistic information when making judgments in the category membership condition but not when making consistency ratings (i.e., the likelihood part of Eq. 1).

In Experiment 2, we explicitly manipulated the strength of the causal relation (the m parameter) following Rehder and Kim (2010; experiment 1) manipulations. Additional to our manipulations in Experiment 1 (rating task condition and coefficient), in Experiment 2, we manipulated information about causal strength (between subjects) by providing information about the probability of the effect given the cause (i.e., $m = 75\%$; $m = 100\%$). At a general level, Bayesian causal models predict that subjects should be sensitive to this information, such that their ratings for each of the four cases (AB, A¬B, ¬AB, ¬A¬B) should change as a function of the causal strength condition in which subjects participated. More specifically, the GM predicts that, if feature base rates are fixed at a value equal to or greater than 0.5 (i.e., features are at least as probable in the category as in other contrasting categories), a stronger causal strength should increase the B regression coefficient and the AB interaction coefficient. The reason for these predictions requires the reader to pay attention to equations in Fig. 1. Because the m parameter appears in equations used to compute the likelihood of cases AB and A¬B, increasing the likelihood for the first equation (the $(m + b - mb)$ term) while decreasing the likelihood for the second one (the $(1 - m)$ term), the overall effect of a larger m parameter will be to increase the contribution of the AB coefficient (i.e., the non-linearity associated with coherence), and at the same time to increase the B coefficient. If subjects in the category membership condition are not sensitive to causal-probabilistic information because they reason by similarity, then they should not show this pattern, whereas it might show in the consistency condition.

3.2.1. Participants

Because in the current experiment we used a third factor, we increased our sample size. Fifty-one participants (32 females), all English speakers aged 18 to 49 (mean = 25.29, *SD* = 6.23) agreed to participate in the online experiment. Three participants were excluded from the analysis (two timed out and one perseverated on the same response for all stimuli). Participants were recruited through the Prolific Academic environment and received monetary compensation according to Prolific rules. All participants were randomly assigned to one of the four experimental conditions (see below) with the constraint that an equal number of participants were in each cell. Same as in Experiment 1, all participants needed to accept the informed consent to continue with the experiment.

3.2.2. Design

In Experiment 2, we set up a 2 (task condition: *consistency*, *category membership*) x 2 (causal strength condition: 75%, 100%) x 3 (regression coefficient: A, B, and AB interaction) mixed design, with the last being a within-subjects factor. Participants learned about a simple causal model (A → B) and then used a rating scale (from 1 to 7) to categorize all possible feature combinations (AB, A¬B, ¬AB, ¬A¬B).

Task conditions (*consistency* and *category membership*) were identical to Experiment 1. In Experiment 2, we added explicit information about causal strength. Subjects received scenarios describing that when feature A was present, it caused $n\%$ of those affected by the category

Table 5
Description of features in Experiment 2 scenario

Feature	*"South American Seasonal Flu of 2018 (SASF-18)"*
A	SP-2 protein in the blood
B	Respiratory difficulty

Table 6
In Experiment 2, raw average ratings for each condition, causal strength and case

		Cases			
Condition	Causal Strength	AB	A¬B	¬AB	¬A¬B
Consistency	75%	6.73 (.47)	4.27 (1.27)	3.73 (2.05)	5.09 (2.12)
	100%	6.67 (.78)	2.50 (1.78)	3.83 (1.64)	4.83 (2.13)
Category	75%	6.42 (1.24)	4.17 (1.34)	3.33 (2.06)	2.50 (1.78)
membership	100%	6.67 (.89)	3.00 (2.30)	2.92 (1.98)	3.25 (2.42)

Note. Standard deviations (*SD*) are in parentheses.

to present feature B, where *n%* could be any of the 75% or 100% explicit probabilities. No explicit probability information about the base rates was provided.

3.2.3. Materials and procedures

Experiment 2 was designed and implemented online through the SurveyMonkey environment. The online booklets and procedures were identical to Experiment 1, with the only difference that we used a novel causal-based scenario and that case order was randomized (i.e., rather than the Latin-square design in Experiment 1). The new scenario described a novel causal model (i.e., *South American Seasonal Flu of 2018, SASF-18*, a type of infection), which again described a simple causal chain relation with one cause and one effect as shown in Table 5.

The "South American Seasonal Flu of 2018 (SASF-18)" scenario reads as follows: "There is a type of infection known as SASF-18 virus (South American Seasonal Flu of 2018). In some cases, this type of infection is characterized by showing the SP-2 protein in blood samples. In some cases, it is also characterized by presenting respiratory difficulties. Not everyone who is infected with the SASF-18 virus shows the SP-2 protein in their blood nor do they have respiratory difficulties. However, when the SP-2 protein is present, it causes 75% of those infected with SASF-18 to present respiratory difficulties because the molecular structure of the SP-2 protein impairs normal pulmonary function." The other causal strength scenario substituted the 75% probability with 100%. The rest of the text remained identical.

3.2.4. Results

To analyze the results of Experiment 2, we followed the same strategy used in Experiment 1. Statistical analyses were performed on the individualized regression coefficients. Table 6 summarizes the raw average ratings for each of the experimental conditions. Consistent with

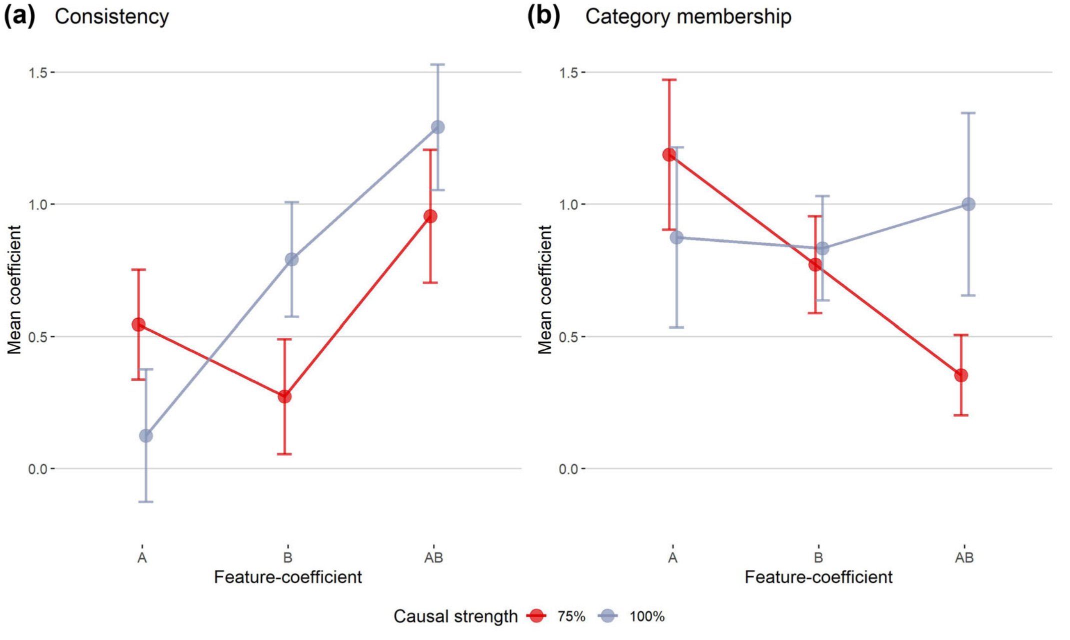

Fig. 3. In Experiment 2, mean regression coefficient weights for each condition.
Note. Mean regression coefficient weights for (a) consistency and (b) category membership task conditions in Experiment 2. Causal strength 100% in gray line and causal strength 75% in orange line.

the GM predictions, we predicted that in the current experiment, the 100% causal strength condition would produce B and AB coefficients that are significantly greater than the A coefficient but that the same pattern would be reduced or not be observed in the 75% causal strength condition. If subjects use causal-probabilistic information when rating category membership, their coefficients should show a similar pattern. Importantly, if subjects in the category membership condition show a different pattern, that would suggest that differences in the relative sizes of coefficients A, B, and AB across task conditions may be related to subjects being less sensitive to causal-probabilistic information in the category membership condition, thus supporting a similarity account of our results.

We submitted individualized regression coefficient data to a 2 (task condition: *consistency*, *category membership*) x 2 (causal strength condition: 75%, 100%) x 3 (regression coefficient: A, B, and AB) mixed ANOVA, with the last being the repeated measures factor. Boxplot analyses using the 1.5 inter-quartile range rule (IQR) revealed one outlier in the consistency and 75% causal strength condition, which we removed from further analysis. By removing this participant, we achieved normality through visual inspection and in the Shapiro–Wilk test for each condition. As Figs. 3 and 4 illustrate, the analysis showed no main effect of condition ($F = 2.3$, $p = .14$), no main effect of causal strength ($F = 1.47$, $p = .23$), and no main effect of coefficient ($F = 0.93$, $p = .40$). We did find a significant two-way interaction between condition and coefficient ($F(2,86) = 4.57$, $MSe = 0.86$, $p = .01$, $\eta_p^2 = 0.10$, power $= 0.76$),

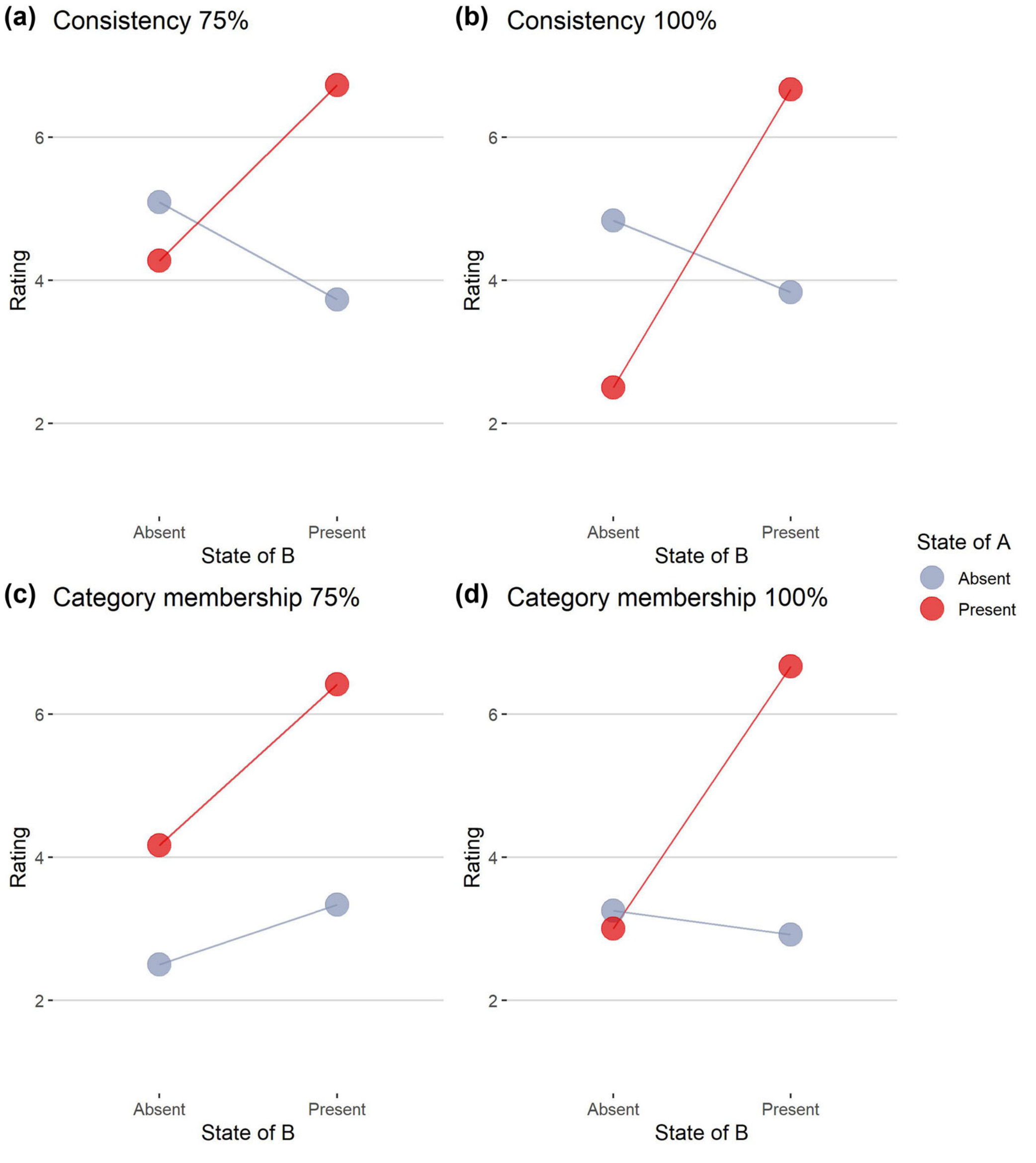

Fig. 4. In Experiment 2, feature interaction plots for each condition and causal strength.
Note. Panels (a) and (b) show interaction plots for consistency condition at 75% (a) and 100% (b) causal strength. Panels (c) and (d) show interaction plots for category membership condition at 75% (c) and 100% (d) causal strength.

a marginally significant interaction between causal strength and coefficient ($F(2,86) = 2.77$, $MSe = 0.86$, $p = .07$, $\eta_p{}^2 = 0.06$, power = 0.53), and a non-significant three-way interaction ($F(2,86) = 0.54$, $p = .59$, $\eta_p{}^2 = 0.01$, power = 0.14). Fig. 3 shows coefficient results for each condition.

We continued the analyses by collapsing across the task condition factor and performing planned comparisons across the repeated measures factor. As the GM predicts, we found a significant interaction between causal strengths (i.e., 75% and 100%) and a planned comparison between the average of B and AB coefficients versus the A coefficient ($F(1,45) = 4.83$, $MSe = 1.45$, $p = .03$, $\eta_p{}^2 = 0.10$, power = 0.58). Further analyses revealed that the interaction occurs because subjects in the causal strength 100% condition weighted more the average of B and AB coefficients than subjects in the causal strength 75% condition ($F(1,45) = 8.28$, $MSe = 0.22$, $p = .006$, $\eta_p{}^2 = 0.16$, power = 0.80) as predicted by the GM.

As Fig. 4 shows, the two task conditions promoted a different pattern when people were rating the ¬A¬B case. To have a test, we followed the two-way interaction between task conditions (i.e., *consistency* and *category membership*) and feature coefficient by collapsing causal strength conditions and testing whether the A, B, and AB coefficients differed across task conditions. As predicted, and as Fig. 3 illustrates, the average of the A and B coefficients was significantly higher in the category membership condition ($F(1,45) = 7.68$, $MSe = 0.36$, $p = .008$, $\eta_p{}^2 = 0.15$, power = 0.77). When comparing the size of the AB interaction term, though the means were in the correct direction according to our hypothesis (i.e., category membership < consistency), the contrast did not achieve significance ($F(1,45) = 2.98$, $MSe = 0.81$, $p = .09$, $\eta_p{}^2 = 0.06$, power = 0.40).

3.2.5. Discussion

Experiment 2 produced even clearer evidence that subjects behave differently when judging category membership than when judging consistency. In Experiment 1, we found evidence when making comparisons across the repeated measures factor. Here, the increase in sample size and consequent increase in power allowed us to find evidence across the between-subjects tasks condition. Subjects in the category membership condition weighted individual features A and B more than subjects in the consistency condition. Though the AB interaction term (reflecting the perceived weight of the causal relation) showed the predicted pattern of means (category membership < consistency), the difference was not significant.

Subjects in this experiment conformed to the GM's predictions regarding the effect that a difference in causal strength (i.e., the *m* parameter) should have. Also, subjects did not show evidence of behaving differently in this regard on both task conditions. This suggests that the hypothesis that subjects in the category membership rating condition were not engaging in causal-probabilistic reasoning, but perhaps in some form of similarity processing, did not receive support. We acknowledge that this conclusion depends in part on the power of the null three-way interaction result, but note that the effect size for the three-way interaction is extremely small, suggesting that the lack of interaction was not due to an underpowered experiment. Experiment 3 will provide more evidence for this conclusion.

3.3. Experiment 3

Experiments 1 and 2 produced evidence suggesting that participants in the category membership condition give more weight to individual features (and less weight to causal dependence information) relative to the consistency condition. This is consistent with the hypothesis that participants in the category membership condition are computing the posterior probability $p(k|o_f)$. According to many theories of categorization, categories are characterized by features that are more probable in category exemplars than in exemplars of contrasting categories (Hampton, 2006; Jones, 1983; Nosofsky, 1992; Rosch & Mervis, 1975; Rosch, Simpson, & Miller, 1976; Tversky, 1977).

To provide more evidence for this account, and to test again the similarity-based alternative explanation, in Experiment 3, we manipulated explicit feature-base rate information. We varied the explicit feature base-rates parameters (i.e., the c and b parameters in Fig. 1 and Table 1). Subjects received a scenario where the A base rate was described as $= 0.7$ and the B base rate $= 0.3$ or a scenario where the A base rate was $= 0.3$ and the B base rate was $= 0.7$ (in fact, we provided this information in percentages, to make base rates easier to understand). Evidently, if subjects reason causal-probabilistically, this manipulation should affect the individual A and B coefficients. We focused our analyses mainly on them. Because we wanted to test our hypotheses with more statistical power than in the previous two experiments, in Experiment 3, we increased the sample size.

3.3.1. Participants

Ninety-six participants (64 females), all English speakers aged 18 to 49 (mean = 30.75, *SD* = 7.51) agreed to participate in the online experiment. The procedures are identical to those described in Experiment 2.

3.3.2. Design

In Experiment 3, we set up a 2 (task condition: *consistency*, *category membership*) x 2 (feature base-rate condition: 30–70, 70–30) x 3 (regression coefficient: A, B, and AB interaction) mixed design, with the last being a within-subjects factor. Similarly to Experiments 1 and 2, participants learned about a simple causal model (A $\rightarrow$ B) and then used a rating scale (from 1 to 7) to judge all possible feature combinations (AB, A¬B, ¬AB, ¬A¬B).

3.3.3. Materials and procedures

As in Experiments 1 and 2, for Experiment 3, we designed online booklets in the SurveyMonkey environment. The online booklets and procedure were identical as in Experiments 1 and 2. In Experiment 3, we used the same scenario as in Experiment 2 (i.e., *"SASF-18"*) apart from omitting causal strength probability information and providing explicit base-rate probabilities. Participants always received the AB case first, with the subsequent order of cases being randomized.

3.3.4. Results

Statistical analyses were performed on the individualized regression coefficients as we did in previous experiments. Raw average ratings are presented in Table 7. We submitted data to a

Table 7
Raw average ratings for each condition, feature base rate and cases in Experiment 3

Condition	Feature Base Rate	AB	A¬B	¬AB	¬A¬B
Consistency	30–70	6.08 (1.18)	2.88 (1.94)	4.13 (1.75)	4.25 (2.29)
	70–30	5.96 (1.16)	3.63 (2.08)	2.71 (1.62)	3.75 (2.21)
Category	30–70	6.35 (.94)	4.13 (1.74)	3.96 (1.58)	2.17 (1.61)
membership	70–30	6.00 (1.31)	4.57 (1.53)	2.57 (1.44)	2.48 (1.76)

Note. Standard deviations (*SD*) are in parentheses.

2 (task condition: *consistency*, *category membership*) x 2 (feature base-rate condition: *30–70*, *70–30*) x 3 (regression coefficient: A, B, and AB) mixed ANOVA, with the last being the repeated measures factor. Following the same IQR rule used in Experiment 2, we removed from further analysis two participants, one in the category membership 30–70 condition and another in category membership 70–30 condition. By removing those participants, we maintain normality assessed through visual inspection and in the Shapiro–Wilk test for three of the four conditions.

ANOVA analyses were corrected using Greenhouse–Geisser. The analysis revealed no main effect of condition ($F = 1.48$, $p = .23$), no main effect of feature base rate ($F = 0.04$, $p = .84$), and a marginally significant main effect of coefficient ($F = 2.78$, $p = .07$). We did find a significant two-way interaction for condition and coefficient ($F(2,180) = 12.49$, $MSe = 0.97$, $p < .001$, $\eta_p{}^2 = 0.12$, power = 0.99), a significant interaction between feature base rate and coefficient ($F(2,180) = 6.66$, $MSe = 0.97$, $p = .002$, $\eta_p{}^2 = 0.07$, power = 0.90), and a non-significant three-way interaction ($F = 0.53$, $p = .58$, $\eta_p{}^2 = 0.006$, power = 0.13). Fig. 5 shows the mean coefficient for each task condition and each feature base rate.

As in Experiment 2, we collapsed across task conditions (i.e., *consistency* and *category membership*) and performed planned comparisons across the repeated measure factor. As expected, we found a significant interaction between base-rate conditions (30–70 and 70–30) and a planned comparison between the A and B coefficients ($F(1,92) = 14.20$, $MSe = 1.66$, $p < .001$, $\eta_p{}^2 = 0.13$, power = 0.96). Overall, subjects in both base-rate conditions were sensitive to probabilistic information. Subjects in condition 30–70 gave less weight to feature A than subjects in condition 70–30 ($F(1,92) = 5.12$, $MSe = 1.01$, $p = -.03$, $\eta_p{}^2 = 0.05$, power = 0.61). Something similar was observed for feature B, where subjects in condition 30–70 gave more weight to feature B than those in condition 70–30 ($F(1,92) = 11.13$, $MSe = 0.60$, $p = .001$, $\eta_p{}^2 = 0.11$, power = 0.91). As shown in Fig. 6, the A and B crossover interaction is evident in both conditions (*consistency* and *category membership*), consistently with the null three-way interaction. From this, we conclude that subjects were sensitive to base-rate information.

Importantly, in Experiment 3, we replicated the two-way interaction between task conditions (i.e., *consistency* and *category membership*) and feature coefficient that was found in Experiments 1 and 2. As shown in Fig. 5, the average of the A and B coefficients was larger in the category membership condition ($F(1,92) = 14.63$, $MSe = 0.335$, $p < .001$, $\eta_p{}^2 = 0.14$, power = 0.97). Conversely, the AB interaction coefficient that reflects the weight of the causal relation in judgments was smaller in the category membership condition ($F(1,92) = 11.13$, $MSe = 0.80$, $p = .001$, $\eta_p{}^2 = 0.11$, power = 0.91).

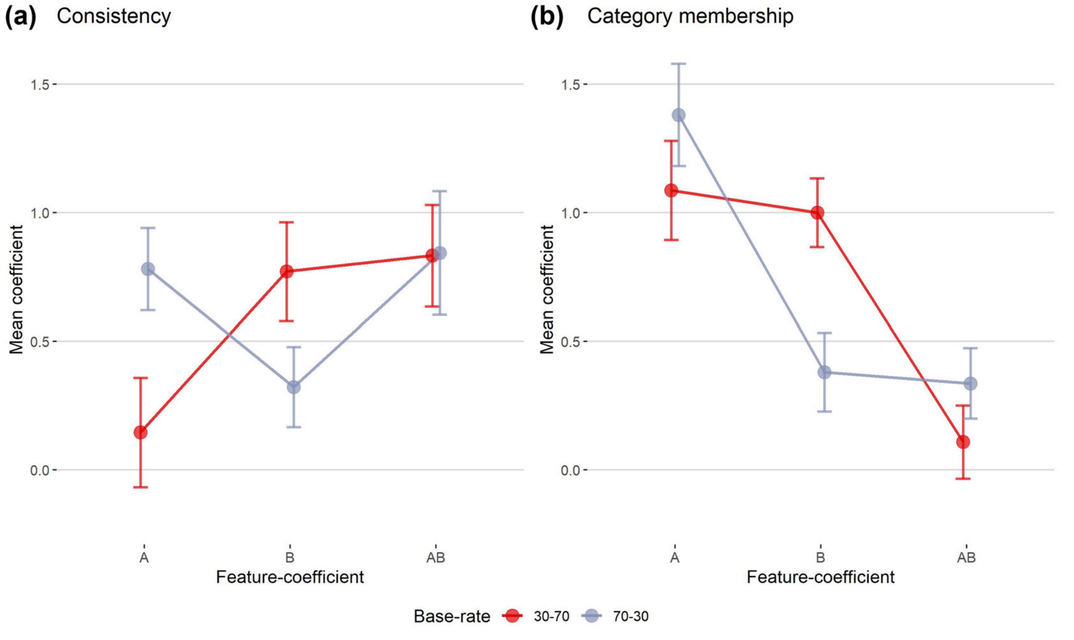

Fig. 5. In Experiment 3, mean regression coefficient weights for each condition.
Note. Mean regression coefficient weights for (a) consistency and (b) category membership task conditions in Experiment 3. Base-rate 70–30 (70% A and 30% B) in gray line and base-rate 30–70 (30% A and 70% B) in orange line.

3.3.5. Discussion

Experiment 3's results strengthen and extend results from the previous two experiments. Subjects in this experiment were appropriately sensitive to base-rate information, contrary to what a similarity account of category membership judgments suggests. Furthermore, Experiment 3 showed even more clearly than Experiments 1 and 2 that subjects in the category membership condition weighted featural information more, and causal information less, compared to the consistency condition. This is what would be expected if participants were computing the posterior probability $p(k|o_f)$ in the category membership condition and assumed that category-specific features are more frequently observed in their category than in alternative categories (e.g., that *barks* is more likely in the *dog* category than in contrasting categories). In the following section, we more formally develop these ideas and show how incorporating them allows a Bayesian model to account for the qualitative pattern of results in our experiments.

4. A Context-dependent causal categorization model

In three studies, we find that people give different patterns of ratings whether they are asked to make consistency judgments or category membership judgments. This result might seem

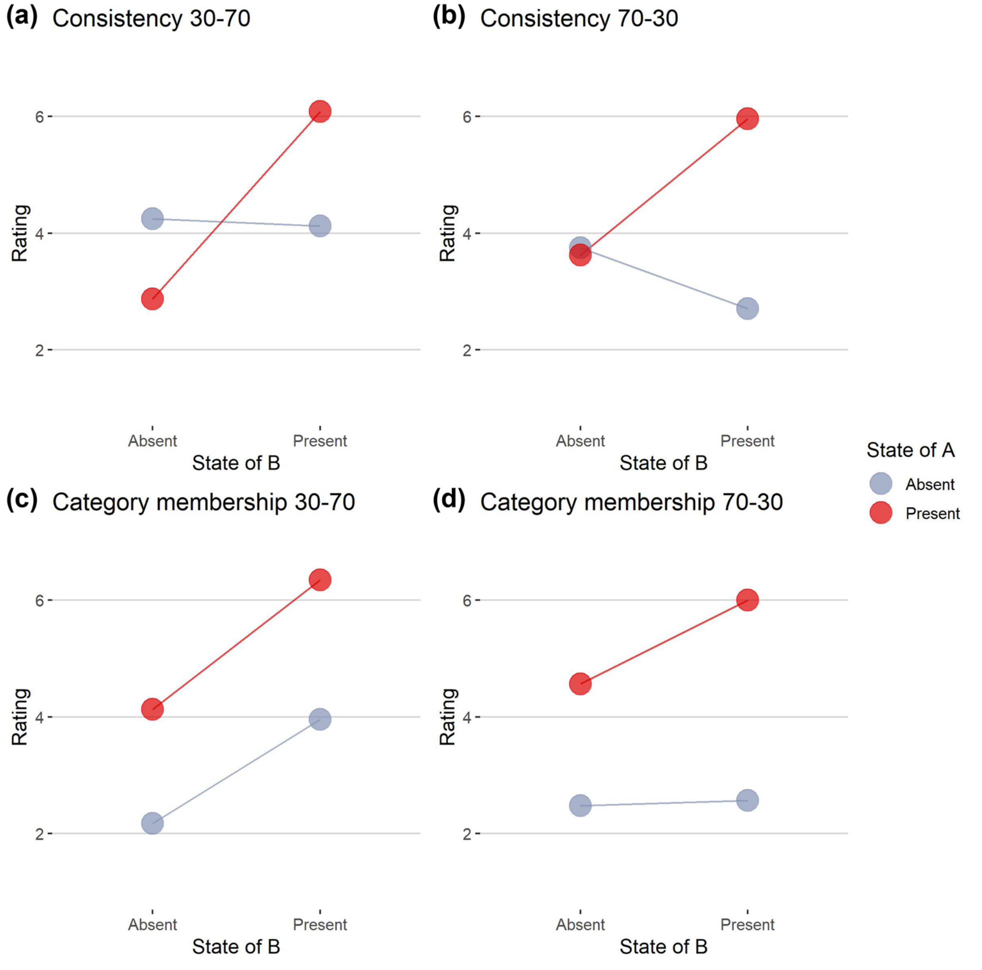

Fig. 6. In Experiment 3, feature interaction plots for each condition and causal strength.
Note: Panels (a) and (b) show interaction plots for consistency condition at base-rate 30–70 (a) and base-rate 70–30 (b). Panels (c) and (d) show interaction plots for category membership condition at base-rate 30–70 (c) and base-rate 70–30 (d).

surprising from the perspective of the GM of causal-based categorization (Rehder, 2003a). According to the model, participants' category membership judgments should track the likelihood that an individual has the observed features, given that the individual belongs to the category that was given. As such, the model seems to predict that participants' category membership judgments should be very similar to their answers to a question that directly elicits likelihood estimates, in contradiction with our results. Alternatively, perhaps participants did not interpret the consistency question as asking them for a likelihood. However, our modeling results shown below suggest that participants' consistency judgments closely track the relevant likelihoods.

We suggest that the inconsistency between the GM and the current results is only apparent. As already discussed in the introductory section, under a Bayesian model of causal-based categorization, we do not always expect people's category membership judgments to track likelihood judgments (for a detailed explanation, see the Supporting Information). Indeed, in our experiments, a Bayesian model makes different predictions whether it assumes that people compute $p(k|o_f)$ or $p(o_f|k)$.

This realization naturally leads to the following hypothesis to explain our data. When participants are asked for consistency judgments, they compute the likelihood $p(o_f|k)$. But when they are asked for category membership judgments, participants compute the posterior probability $p(k|o_f)$, which depends in part on the probability of observing *f* in an object that does not belong to category *k*.

Below, we formalize this hypothesis, and we explain why, in conjunction with the assumption that features A and B are more prevalent in the focal category than in alternative categories, it can account for our empirical results. Then we quantitatively assess the fit of our model to the data.

4.1. Posterior probability in causal-based categorization

Computing the full posterior probability $p(k|o_f)$ (rather than simply the likelihood) introduces some complications. Remember that following Bayes' rule (Eq. 1), $p(k|o_f)$ depends in part on the probability of observing an object with features *f*, $p(o_f)$. This quantity can be computed as

$$p\left(o_f\right) = p\left(o_f|k\right) p\left(k\right) + p\left(o_f|\neg k\right) p\left(\neg k\right), \tag{2}$$

where $p(o_f)$ depends in part on the likelihood $p(o_f|\neg k)$ of observing o_f assuming that the object does not belong to *k*.

Just like the likelihood $p(o_f|k)$ depends on the observer's causal model of category *k*, the likelihood $p(\neg k)$ depends on the observer's causal model of all other possible categories that the object could belong to. To compute this likelihood, one could in principle enumerate all possible categories other than *k* and define a causal model for each of them. However, it is implausible to assume that participants explicitly represent every other category that the object might fall under, let alone have a dedicated causal model for each of them. Instead, we can assume that participants use a "catch-all" causal model for $\neg k$, which parallels the one they use for category *k*. That is, they have a causal model that defines the average base rate of features A and B, as well as the average causal strength of A for B, in members of categories other than *k*. Following the GM nomenclature, we call these parameters *qc*, *qb*, and *qm*; see Fig. 7.

Once these parameters are set, we can compute $p(o_f|\neg k)$ using the likelihood equations given earlier (see Table 1) for $p(o_f|k)$, substituting *qc*, *qm*, and *qb* for *c*, *m*, *b*.

4.2. Explaining differences in the importance of individual features across conditions

In line with many analyses of categorization (Hampton, 2006; Jones, 1983; Rosch & Mervis, 1975), we assume that people hold that features A and B are rare outside of

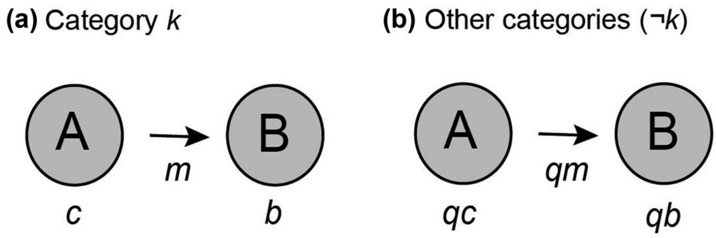

Fig. 7. Context-dependent Bayesian model parameters.
Note. The observer has a causal model for category k (a), and a "catch-all" causal model for all other categories (b). qc = A's base rate, qb = B's base rate, qm = causal parameter corresponding to the probability of the effect given the cause.

category k. For example, in one of our scenarios, we told participants that people with the *SASF* disease often have protein SP-2 in their blood, and from this information, it seems plausible that healthy people rarely have protein SP-2 in their blood. This assumption is also consistent with the idea that features are sparse in general (for any given feature, most things do not possess that feature), an idea that is consistent with findings from other areas of human reasoning (Hendrickson, Navarro, & Perfors, 2016; Klayman & Ha, 1987; Oaksford & Chater, 1994; see also Navarro & Perfors, 2011).

Given this assumption, the presence or absence of individual features plays a much larger role in posterior probability estimation relative to likelihood estimation. The main effect of the presence versus absence of feature f_i on likelihood estimation can be computed as $p(f_i \mid k)$-$p(\neg f_i \mid k)$, which means that the presence of feature f_i has an effect on likelihood estimation to the extent that $p(f_i \mid k) > .5$. By contrast, the main effect of feature f_i on posterior probability estimation is a function of $p(f_i \mid k)/p(f_i \mid \neg k)$. Assuming that f_i is more prevalent in category k than in other categories, f_i is expected to have a large main effect on posterior probability estimates.

For instance, suppose that feature A is present in about 50% of members of category k and present in only 10% of members of other categories. The presence versus absence of A is not expected to have a main effect on likelihood estimates (since $p\,(A \mid k) = p(\neg A \mid k)$. By contrast, we expect A to have a large effect on posterior probability estimates since $p(A \mid k) > p(A \mid \neg k)$.

Hence, while coherence matters for both likelihood and posterior probabilities, its role is more apparent for likelihood estimates, where individual features play a comparatively smaller role. This is precisely the qualitative pattern we find across our three experiments.

We also note that, while we find a decreased coherence effect in the category membership condition, the effect is still present. To explain why the effect still occurs, we need to assume $qm < m$: Participants think that the causal strength of the relationship between A and B is likely to be higher inside category k than outside. If participants did not assume such a difference, then coherence information would not be diagnostic about category membership: With $qm = m$, we expect the same degree of covariation between A and B regardless of whether the item belongs to category k or not. Analyses in the Supporting Information confirm that our account gives a better account of the data when $qm < m$.

4.3. Modeling approach

We implement our hypothesis as a "context-dependent" model that assumes that people compute the likelihood $p(o_f|k)$ when they make consistency judgments and that they compute the posterior probability $p(k|o_f)$ when they make category membership judgments. The reason for this difference is, as previously discussed, that when categorizing, subjects seem to make assumptions about alternative categories but not when making consistency judgments. We also consider a model that assumes that people always compute likelihoods (this is equivalent to the standard formulation of the GM) and a model that assumes that people always compute posterior probabilities. Each model has free parameters representing the parameter values of the causal model for category k: c, m, and b. We assume that these parameter values are the same in the consistency and the category membership conditions—that is, the question we ask participants does not affect their causal models. This assumption means that none of the free parameters in the model is specific to a particular condition, so the model cannot account for the differences between category membership and consistency judgments simply by tweaking these parameters.

We set the values of the "other-categories" causal model parameters as $qc = 0.1$, $qm = 0.2$, and $qb = 0.1$. We also set the base rate of category k to a low value, $p(k) = 0.2$. These values are to some extent arbitrary, but they reflect the hypothesis (discussed earlier) that participants assume that category-specific features are in general rare outside of the target category (and similarly, that categories have relatively few members).[1] In the Supporting Information, we show that our results are robust across a wide range of possible values for qc, qm, qb, and $p(k)$ as long as qc is not too high.

Because Experiment 1 has a different cover story than Experiments 2 and 3, but Experiments 2 and 3 share the same cover story, we analyzed data from Experiments 2 and 3 together and data from Experiment 1 separately.

Following Rehder (2015), we map model judgments to Likert ratings using the formula:

$$Rating = 6p^{\gamma} + 1, \tag{3}$$

where γ is a free parameter.[2] We then fit the model by finding the values of c, m, b, and γ that jointly minimize the root mean squared error (RMSE) between mapped model predictions and average participant judgments, using the optim function in R (R core team, 2022).

4.4. Modeling results

Like participants, the model exhibits a stronger coherence effect in the consistency conditions than the category membership conditions, see Figs. 8 and 9. Coherence with the causal relationship has a large influence on the model's likelihood judgments but a somewhat smaller influence on the model's posterior probability judgments.

4.4.1. Modeling results for Experiment 1

We find that the context-dependent model accounts for the data well (RMSE = 0.449), slightly better than the likelihood and the posterior models (RMSE = 0.473, for both models). However, due to the small number of data points in this preliminary experiment (eight conditions in total), we find that the context-dependent model also easily overfits the data,

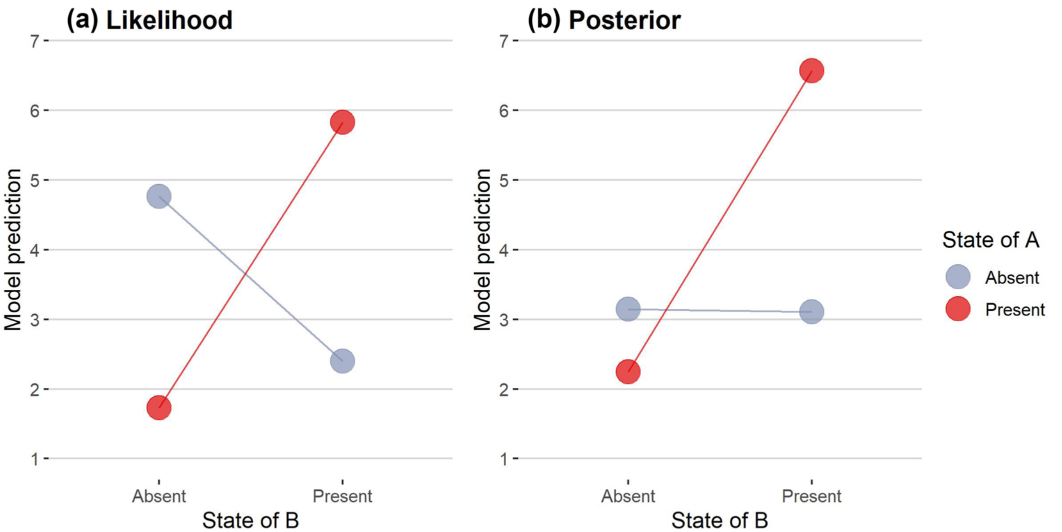

Fig. 8. Model predictions for Experiment 1.

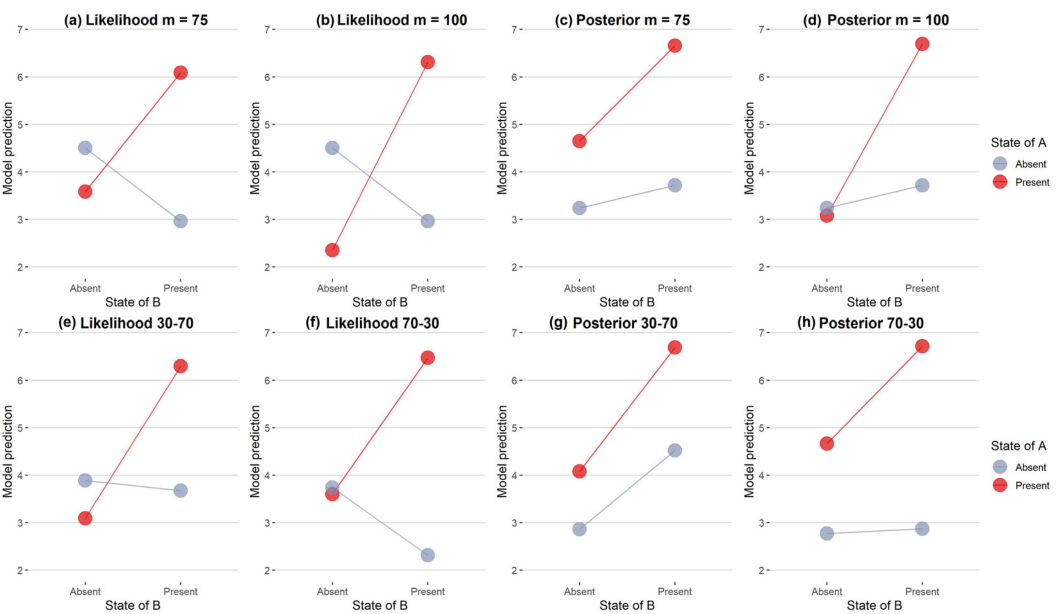

Fig. 9. Model predictions in Experiments 2 and 3.
Note. Upper panels present the model's predictions for each of the between-subjects conditions in Experiment 2. Bottom panels show the model's predictions for conditions used in Experiment 3.

Table 8
Best-fitting parameter values for the context-dependent model in Experiments 2 and 3

	c	m	b	γ
Low	0.76 *(Experiment 3, $c = 0.3$)*	0.84 *(Experiment 2, $m = 0.75$)*	0.12 *(Experiment 3, $b = 0.3$)*	0.36
Unspecified	0.73	0.87	0.17	
High	0.87 *(Experiment 3, $c = 0.7$)*	0.97 *(Experiment 2, $m = 1$)*	0.45 *(Experiment 3, $b = 0.7$)*	

Note. When relevant, the value explicitly communicated to participants is in parentheses. Explicit values in the instructions were provided for m in Experiment 2 and for c and b in Experiment 3.

which makes it perform less well under cross-validation. When using leave-one-out cross-validation,[3] the model performs less well (RMSE = 1.112) than the likelihood model (RMSE = 0.945) and the posterior model (RMSE = 0.943). Therefore Experiment 1 is not on its own conclusive regarding our hypothesis.

We also computed model fit on the non-aggregated data, using the Bayesian information criterion (BIC)[4]. The context-dependent model had a better fit to the data (BIC = 333.1) relative to the likelihood model (BIC = 339.4; BF = 23.29) and the posterior model (BIC = 338.1; BF = 11.7). It also outperformed a random baseline (BIC = 373.6; BF $> 10^4$) and a saturated model (BIC = 438.2; BF $> 10^4$).[5]

4.4.2. Modeling results for Experiments 2 and 3

The predictions of the context-dependent model were highly correlated with participants' consistency judgments, $r(14) = .96$, $p < .001$, and with their category membership judgments, $r(14) = .98$, $p < .001$, see Fig. 10.

The context-dependent model had a better fit to the data (RMSE = 0.465) than the posterior model (RMSE = 0.589) or the likelihood model (RMSE = 0.574). This pattern held even when we computed model fit using leave-one-out cross-validation (context-dependent model: RMSE = 0.693, posterior model: RMSE = 0.896, likelihood model: RMSE = 0.859).

In Experiments 2 and 3, we explicitly provided information to participants about causal strength (Experiment 2) or category-specific feature base rates (Experiment 3). Because participants might not have used the exact values we gave them when completing the task, we fit the values of these parameters to the data, using a different free parameter for each condition. For instance, in Experiment 2, participants in the low causal strength condition were told that $m = 0.75$, and participants in the high causal strength condition were told that $m = 1$, so we use one free parameter for each condition. This strategy allows us to assess the extent to which the manipulations succeeded in influencing participants' representations.

Recovered parameter values suggest that the manipulations, in both studies, were largely successful (see Table 8). In Experiment 2, the best-fitting value of m is higher in the high causal strength condition, compared to the low causal strength condition, and the values obtained are relatively close to those communicated to participants. In Experiment 3, the best-fitting value of c is higher in the *high-c* compared to the *low-c* condition, and a similar

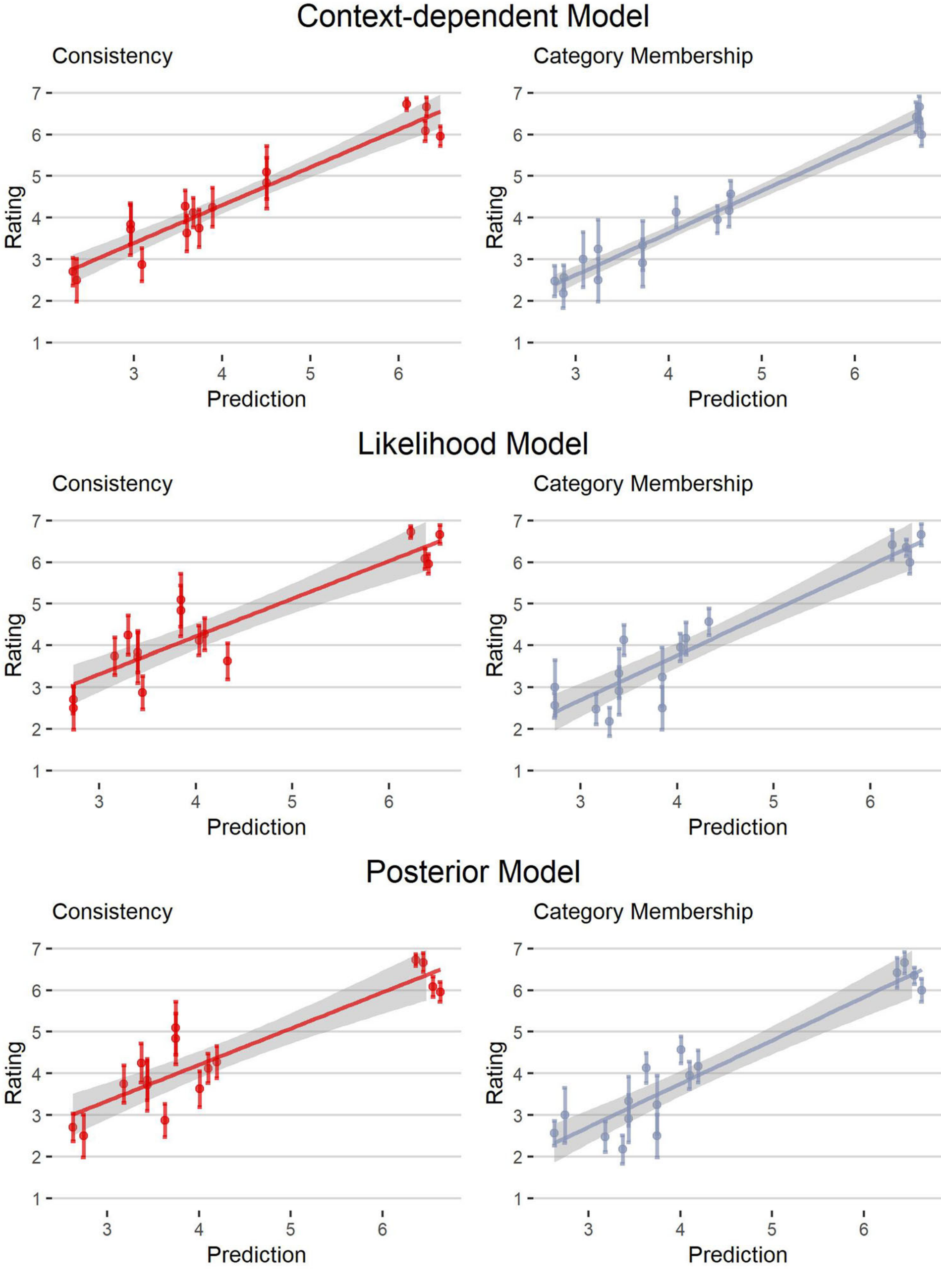

Fig. 10. Average participant ratings as a function of model predictions.
Note. Average participant ratings as a function of model predictions in Experiments 2 and 3 for our context-dependent model (top), the likelihood model (middle), and the posterior model (bottom). Error bars display the standard error of the mean.

pattern obtains for parameter b. At the same time, parameter values in Experiment 3 reveal that participants might have been biased toward higher values of c and lower values of b, compared to those they were given. One speculative explanation is that cause features are rarely less prevalent than effect features (there are more birds that have wings than birds who fly; Ahn & Kim, 2001; Rehder & Kim, 2010), and participants might have "corrected" the numbers they were given to reflect this. In the Supporting Information, we show that our results are robust across a wide range of possible values for qc, qm, qb (the causal model parameters for $\neg k$) and $p(k)$ as long as qc is not too high.

We also computed model fit on the non-aggregated data, using the BIC. The context-dependent model had a better fit to the data (BIC = 2053.3) relative to the likelihood model (BIC = 2073.1; BF $> 10^4$) and the posterior model (BIC = 2075.6; BF $> 10^4$). It also outperformed a random baseline (BIC = 2195.0; BF $> 10^4$) and a saturated model (BIC = 3573.0; BF $> 10^4$).

4.5. Modeling discussion

Our findings suggest that participants interpreted the consistency question as asking them for a likelihood (the probability of observing the current features, given membership in the category), while they interpreted the category membership question as asking them for a posterior probability (the probability of belonging to the category, given the observed features). This hypothesis was more consistent with the data than the hypothesis that participants interpreted both questions as asking for a likelihood or interpreted both questions as asking for a posterior probability. The context-dependent model is able to account for the results obtained in our three experiments.

The current results are quite consistent with the spirit of the GM. We find that we are able to closely predict people's judgments (both about consistency and category memberships) by assuming that they make probabilistic computations over a generative causal model of the category. On the other hand, our results also suggest that the GM will not always give accurate approximations of people's judgments by simply assuming that they compute a likelihood for the received causal model. Our explanation for this context dependency is that, by virtue of being confronted with a categorization task, people bring to mind assumptions relative to category features being of high probability within the category and of low probability in alternative categories, and thus likelihood and posteriors systematically diverge.

5. General discussion

In three causal-based categorization experiments, we asked participants to either rate whether an exemplar belonged to a category or whether the exemplar would be expected given the category. Across our three experiments, subjects consistently treated the two dependent variables differently. Overall, our regression coefficients analysis revealed that the category membership condition induced subjects to weight the individual features more and the causal relation less than in the consistency condition. Another way of describing these differences is that across our three experiments, the coherence effect was larger for consistency than for category membership judgments. Participants' judgments in both conditions were also sensitive to manipulations of parameters of the category's causal model, suggesting that they did engage in causal-probabilistic computations on both conditions and did not simply respond as a function of similarity.

We formally accounted for these effects by assuming that in the consistency condition, people estimate the likelihood of observing a feature pattern given the category, while in the category membership condition, they estimate the posterior probability of the category given the observed feature pattern. Our context-dependent model correctly captured the qualitative patterns we observed in our data and achieved a close numerical fit to the data. Subjects' ratings in the consistency condition were predicted by assuming that subjects were tracking exclusively the feature patterns' likelihood (i.e., $p(o_f|k)$). In contrast, subjects' ratings in the category membership condition were predicted by assuming that subjects were tracking the full posterior probabilities (i.e., $p(k|o_f)$). Overall, our results suggest that people engage in causal-based categorization but that their judgements are context-dependent and flexibly adapt to the question they are being asked.

Our context-dependent model is an extension of a previous Bayesian model for causal categorization, the GM (Rehder, 2003a, 2003b, 2015). The GM assumes that categories are generated by the causal relation of their features, thus certain feature combinations are more likely to be generated given the queried category's causal model. The GM assumes that people are only computing likelihood probabilities in categorization experiments. In consequence, the GM stipulates a series of likelihood equations (presented in Table 1) for each feature combination. The likelihood model we use to predict consistency judgments is in fact equivalent to the GM model when a single category is used. However, our behavioral results suggest that people are not always estimating likelihood when categorizing. We found a relatively bad fit of the likelihood model for the category membership condition. In contrast, the full posterior model does a better job of capturing the empirical pattern observed in category membership ratings. A notable characteristic of our model is the use of a "catch-all" causal model for all the other possible categories (see Fig. 7), intended to capture background assumptions about the task itself (i.e., sparsity).

5.1. Expanding on the sparsity assumption

Our model relies on a sparsity assumption (Hendrickson et al., 2016; Klayman & Ha, 1987; Navarro & Perfors, 2011; Oaksford & Chater, 1994): When categorizing, people assume that category-typical features have low prevalence outside the focal category. In particular, our analyses suggest that people assume that causal strength should be relatively lower outside than inside the focal category (i.e., $qm < m$) and also assume that the causal feature's prevalence is higher in the focal category than in alternative categories (i.e., $c > qc$; see Supporting Information II). The sparsity assumption is not exclusive to our causal-based categorization model and is important in other domains of cognition. Oaksford and Chater (1994) showed that a Bayesian reasoner is likely to assume that antecedent and consequent features in a conditional rule (i.e., if p then q) are sparse. The same assumption has also been shown to be critical for syllogistic reasoning (Chater & Oaksford, 1999), dual-factor heuristics in causal judgment (Hattori & Oaksford, 2007), and hypothesis testing (Hendrickson et al., 2016). A critical assumption for explaining our results is $c > qc$: The feature that is causally "deep" in the category's causal model (feature A in a causal model A $\rightarrow$ B) is assumed by participants to be more prevalent inside than outside the category.

5.2. Category coherence

Many studies in the literature on concepts and categorization have shown that whenever a certain combination of features is expected for a given category, people judge exemplars that preserve that relationship to be better category exemplars than those that do not (Malt & Smith, 1984; Murphy & Wisniewski, 1989; Rehder, 2003a, 2003b,2015; Rehder & Kim, 2006, 2010; Rehder & Ross, 2001; Wisniewski, 1995; see also Rehder & Hastie, 2004). This is known as the category coherence effect. Many concepts in everyday categorization including natural kinds exhibit a coherence effect (Hampton et al., 2009; Malt & Smith, 1984), though there is evidence suggestive that artifacts may not (Puebla & Chaigneau, 2014). Our behavioral experiments and computational modeling are largely consistent with the previous literature.

Our work also extends the literature on the coherence effect by demonstrating its existence in a task where participants are asked to rate likelihoods—whether a particular combination of features is likely to be observed, given that the item belongs to the focal category. Furthermore, we find that the magnitude of the effect is higher in this task than in a task involving category membership judgments, although the effect is still present in the latter type of judgments (see also Marchant & Chaigneau, 2020). We account for this difference via a rational analysis of the tasks, which highlights the fact that category membership judgments rely on the assumptions people make about the prevalence of features across categories. Our analysis assumes that people use a stable causal model to reason about a category but use this model flexibly as a function of the probabilistic computations that are implied by a given query.

In the psychology of concepts, coherence effects have been found in tasks that are closely related—but not identical—to category membership judgments. For example, Rehder and Hastie (2004) found a coherence effect in a property induction task. In such a task, participants are shown an individual that has feature C and are asked to what extent members of category k are likely to have feature C. Participants are more likely to make this generalization if the other features of the focal individual are coherent under category k (e.g., *k's* causal model is A $\rightarrow$ B and the individual has both A and B). Such findings indicate that the GM framework can explain human behavior in many different tasks. Extrapolating from the present findings, we suggest that many of these tasks may also involve posterior probability computations. For example, if an individual has a combination of features that makes it likely to belong to category k, people might be willing to extend properties of the individual to other members of k because they are confident that the individual is a true member of k. Ultimately, predicting the kinds of computations that people engage in when they solve a given categorization problem is likely to benefit from a careful task analysis of the problem.

5.3. Future experiments

Our Bayesian analysis of causal-based categorization generates new predictions to be explored in future work. In existing work, researchers have manipulated the parameters of a category's causal model and shown that these manipulations affect people's categorization judgments. Our analysis implies that manipulating parameters *outside* of that causal model would also affect participants' judgments. That is, participants' intuitions about category membership should be sensitive to the prevalence of category-typical features (and the causal relationship between them) in non-category members.

Because we explain our novel effect (a difference in the role of coherence in consistency and category membership judgments) with a sparsity assumption, one should be able to diminish or abolish the effect by giving participants instructions that are inconsistent with sparsity. In our analysis, coherence will matter for category membership to the extent that the causal relationship between A and B is stronger inside than outside the focal category. In the limit, when A causes B to the same extent for category members and non-members, the coherence effect should be very weak or absent. Similarly, if sparsity were contradicted and participants were taught that the A cause is as prevalent outside as it is inside the focal category, then our model predicts that coherence should increase for category membership judgments. Intuitively, this would occur because the causal feature becomes non-predictive of the category, and the causal relationship becomes the only information predictive of the category.

Finally, our findings suggest that people are able to flexibly adjust the computations they make depending on the query. However, it remains unknown how many people are able to do so. In the studies we report here, we presented consistency and category membership questions to different groups of participants. As such, we do not yet know how many participants are effectively disposed to give different answers to our different queries. Conducting within-subjects tests could shed some light on this issue.

6. Conclusion

Our work provides further evidence in support of a generative Bayesian framework for causal-based categorization. Our model adds to the literature the idea that people make different judgments depending on whether the task suggests they should compute a likelihood or a posterior probability and that posteriors are more appropriate to capture behavioral patterns in categorization. Consistent with this prediction, we found that people were less sensitive to causal coherence when making judgments of category membership, compared to when they were assessing consistency. Additionally, the model was able to capture the way in which manipulations of causal and probabilistic information affected participants' judgments, showing that in all our experiments and conditions, participants were in fact doing causal-based reasoning.

Acknowledgments

This research was supported by ANID Fondecyt grant No. 1190006 to the third author and a doctoral scholarship from Universidad Adolfo Ibáñez to the first author. We gratefully acknowledge thoughtful reviews by Bob Rehder, Mike Oaksford, and an anonymous reviewer. We are also grateful to Neil Bramley, Chris Lucas, and members of their labs at the University of Edinburgh, for an insightful discussion about this work.

Open Research Badges

This article has earned Open Data badges. Data are available at osf.io/qh23f/?view_only=d948400f5c69451ebf15a12f87d3169e.

Notes

1 We note that two other particularly salient possible values for *qm* are *qm=0* (no causal relationship between A and B outside the category) or *qm=m* (the causal relationship between A and B does not depend on the category). A Bayesian reasoner who holds both possibilities as likely should assign an expected value to *qm* that is higher than 0 but lower than *m*, motivating our choice of an intermediate value for *qm*.

2 To limit the number of free parameters, we use the same γ parameter for both the consistency and the category membership condition—effectively assuming that the mapping between participants' internal representations and their Likert ratings is similar for both types of judgments. We note that this choice works *against* our hypothesis: the context-dependent model assumes that participants are computing different types of representations when asked to make consistency versus category membership judgments, which suggests that people may actually use a different response function when making these different types of judgments.

3 We perform leave-one-out cross-validation by removing data from one condition (e.g., consistency judgments for the A∼B observation) and training the model on the remaining data. Then we compute the squared error between model prediction and average human judgment for the removed condition (that the model did not "see" during training). We repeat the procedure for each condition to compute an RMSE error score.

4 Lower BIC values indicate better fit. Computing the BIC requires making assumptions about the probability that a participant would make a given Likert rating, given a model prediction. Here, we modeled participants' ratings as samples from a truncated-discretized normal distribution with mean m and variance σ^2, where m is the model prediction. We fit the parameter σ at the group level. We use the same approach to computing BIC in studies 2-3.

5 Bayes factors are derived from the BICs of the respective models using the formula $BF = exp(-\frac{\Delta BIC}{2})$.

References

Ahn, W., & Kim, N. S. (2001). The causal status effect in categorization: An overview. *Psychology of Learning and Motivation*, *40*, 23–65. https://doi.org/10.1016/s0079-7421(00)80017-1

Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology*, *41*(4), 361–416. https://doi.org/10.1006/cogp.2000.0741

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–249.

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, *30*(3), 241–254. https://doi.org/10.1017/S0140525x07001653

Chater, N., & Oaksford, M. (1999). The Probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, *38*(2), 191–258. https://doi.org/10.1006/cogp.1998.0696

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

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*(3), 227–247. https://doi.org/10.1037/0096-3445.117.3.227

Gluck, M. A., & Myers, C. (2001). *Gateway to memory: An introduction to neural network modeling of the hippocampus and learning*. Cambridge, MA: MIT Press.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085–1108. https://doi.org/10.1037/a0028044

Hagmayer, Y. (2016). Causal Bayes nets as psychological theories of causal reasoning: Evidence from psychological research. *Synthese*, *193*(4), 1107–1126. https://doi.org/10.1007/s11229-015-0734-0

Hampton, J. A. (2006). Concepts as prototypes. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 79–113). Amsterdam, the Netherlands: Elsevier Academic Press. https://doi.org/10.1016/S0079-7421(06)46003-5

Hampton, J. A., Storms, G., Simmons, C. L., & Heussen, D. (2009). Feature integration in natural language concepts. *Memory & Cognition*, *37*(8), 1150–1163. https://doi.org/10.3758/MC.37.8.1150

Hattori, M., & Oaksford, M. (2007). Adaptive non-interventional heuristics for covariation detection in causal induction: Model comparison and rational analysis. *Cognitive Science*, *31*(5), 765–814. https://doi.org/10.1080/03640210701530755

Hendrickson, A. T., Navarro, D. J., & Perfors, A. (2016). Sensitivity to hypothesis size during information search. *Decision*, *3*(1), 62–80. https://doi.org/10.1037/dec0000039

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology*, *62*, 135–163. https://doi.org/10.1146/annurev.psych.121208.131634

Jones, G. V. (1983). Identifying basic categories. *Psychological Bulletin*, *94*(3), 423–428. https://doi.org/10.1037/0033-2909.94.3.423

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211.

Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior*, *23*(2), 250–269. https://doi.org/10.1016/S0022-5371(84)90170-1

Marchant, N., Canessa, E., & Chaigneau, S. E. (2022). An adaptive linear filter model of procedural category learning. *Cognitive Processing*, *23*, 393–405. https://doi.org/10.1007/s10339-022-01094-1

Marchant, N., & Chaigneau, S. E. (2020). Modulating the coherence effect in causal-based processing. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 2531–2527). Toronto: Cognitive Science Society.

Marsh, J. K., & Ahn, W. K. (2006). Order effects in contingency learning: The role of task complexity. *Memory & Cognition*, *34*(3), 568–576. https://doi.org/10.3758/BF03193580

Murphy, G. L., & Wisniewski, E. J. (1989). Categorizing objects in isolation and in scenes: What a superordinate is good for. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 572–586. https://doi.org/10.1037/0278-7393.15.4.572

Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, *118*(1), 120–134. https://doi.org/10.1037/a0021110

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *Essays in honor of William K. Estes, Vol. 1. From learning theory to connectionist theory; Vol. 2. From learning processes to cognitive processes* (pp. 149–167). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608–631. https://doi.org/10.1037/0033-295X.101.4.608

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, UK: Cambridge University Press.

Puebla, G., & Chaigneau, S. E. (2014). Inference and coherence in causal-based artifact categorization. *Cognition*, *130*(1), 50–65. https://doi.org/10.1016/j.cognition.2013.10.001

Quillien, T. (2018). Psychological essentialism from first principles. *Evolution and Human Behavior*, *39*(6), 692–699. https://doi.org/10.1016/j.evolhumbehav.2018.07.003

R Core Team. (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/

Rehder, B. (2003a). Categorization as causal reasoning. *Cognitive Science*, *27*(5), 709–748. https://doi.org/10.1016/S0364-0213(03)00068-5

Rehder, B. (2003b). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning Memory and Cognition*, *29*(6), 1141–1159. https://doi.org/10.1037/0278-7393.29.6.1141

Rehder, B. (2010). Causal-based categorization. *Psychology of Learning and Motivation*, *52*, 39–116. https://doi.org/10.1016/s0079-7421(10)52002-4

Rehder, B. (2015). Independence and dependence in human causal reasoning. *Cognitive Psychology*, *72*, 54–107. https://doi.org/10.1016/j.cogpsych.2014.02.002

Rehder, B. (2017). Concepts as causal models: Categorization. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 347–376). New York, NY: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199399550.013.39

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*(3), 323–360. https://doi.org/10.1037/0096-3445.130.3.323

Rehder, B., & Hastie, R. (2004). Category coherence and category-based property induction. *Cognition*, *91*(2), 113–153.

Rehder, B., & Kim, S. (2006). How causal knowledge affects classification: A generative theory of categorization. *Journal of Experimental Psychology: Learning Memory and Cognition*, *32*(4), 659–683. https://doi.org/10.1037/0278-7393.32.4.659

Rehder, B., & Kim, S. W. (2010). Causal status and coherence in causal-based categorization. *Journal of Experimental Psychology: Learning Memory and Cognition*, *36*(5), 1171–1206. https://doi.org/10.1037/a0019765

Rehder, B., & Ross, B. H. (2001). Abstract coherent categories. *Journal of Experimental Psychology: Learning Memory and Cognition*, *27*(5), 1261–1275. https://doi.org/10.1037/0278-7393.27.5.1261

Rips, L. J. (2008). Causal thinking. In J. E. Adler & L. J. Rips (Eds.), *Reasoning: Studies of human inference and its foundations* (pp. 597–631). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO97805118114273.031

Rosch, E., & Mervis, C. B. (1975). Family resemblance studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605. http://doi.org/10.1016/0010-0285(75)90024-9

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human perception and performance*, *2*(4), 491–502. https://doi.org/10.1037/0096-1523.2.4.491

Rottman, B. M., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin*, *140*(1), 109–139. https://doi.org/10.1037/a0031903

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology*, *66*, 223–247. https://doi.org/10.1146/annurev-psych-010814-015135

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352. https://doi.org/10.1037/0033-295X.84.4.327

Waldmann, M. R., & Hagmayer, Y. (2006). Categories and causality: The neglected direction. *Cognitive Psychology*, *53*(1), 27–58. https://doi.org/10.1016/j.cogpsych.2006.01.001

Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal models and the acquisition of category structure. *Journal of Experimental Psychology: General*, *124*(2), 181–206. https://doi.org/10.1037/0096-3445.124.2.181

Wisniewski, E. J. (1995). Prior knowledge and functionally relevant features in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 449–468. https://doi.org/10.1037/0278-7393.21.2.449

Zhao, B., Lucas, C. G., & Bramley, N. R. (2021). How do people generalize causal relations over objects? A non-parametric Bayesian account. *Computational Brain & Behavior*, *5*, 22–44. https://doi.org/10.1007/s42113-021-00124-z

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1. Error for each model as a function of the "other-categories" causal model parameters.

Fig. S2. Model fit for each model as a function of the "other-categories" causal model parameters.