Causal Status meets Coherence: The Explanatory Role of Causal Models in Categorization

Ralf Mayrhofer (rmayrho@uni-goettingen.de)
Anselm Rothe (anselm.rothe@stud.uni-goettingen.de)
Department of Psychology, University of Göttingen,
Goßlerstraße 14, 37073 Göttingen, Germany

Abstract

Research on causal-based categorization has found two competing effects: According to the causal-status hypothesis, people consider causally central features more than less central ones. In contrast, people often focus upon feature patterns that are coherent with the category's causal model (coherence hypothesis). Following up on the proposal that categorization can be seen as inference to the best explanation (e.g., Murphy & Medin, 1985), we propose that causal models might serve different explanatory roles. First, a causal model can serve as an explanation why the prototype of a category is as it is. Second, a causal model can also serve as an explanation why an exemplar might deviate from the prototype. In an experiment, we manipulated whether typical or atypical features were linked by causal mechanism. We found a causal-status effect in the first case and a coherence effect in the latter one, suggesting both are faces of the same coin.

Keywords: categorization; causal reasoning; causal status effect; coherence effect; explanation.

Introduction

The question how people organize objects into categories and form abstract concepts about the world to make sense of it has puzzled philosophers for centuries. It is therefore not surprising that categorization has been an important topic in cognitive science since its beginnings. Early but nevertheless prominent accounts concentrated on the role of similarity between exemplars, or exemplars and category prototypes, or rules with respect to defining features of a category (e.g., Nosofsky, 1986; Rosch & Mervis, 1975; for an overview see Ashby & Maddox, 2005). In contrast, more recent accounts emphasize the role of abstract conceptual, mostly causal knowledge as an integral part of category representations (see Murphy & Medin, 1985; Rehder, 2010; Rehder & Hastie, 2001; Sloman, Love, & Ahn, 1998): People do not only know which features are typical for a category and which not. They often represent knowledge about how strongly and *why* features are correlated with each other within a category (Ahn, Marsh, Luhmann, & Lee, 2002; Murphy & Medin, 1985). For instance, people do not only know that birds typically have wings, can fly, and build nests on trees. People also know that birds build nests on trees because they can fly and that they can fly because they have wings.

This kind of causal knowledge underlying category concepts can be formalized in causal graphical models or Bayes nets (see Rehder, 2003a, 2003b; Waldmann, Holyoak, & Fratianne, 1995). A causal Bayes net consists of nodes, which represent causally relevant variables (i.e., in case of categorization: the presence or absence of features or—more general—properties of objects), and arrows, which stand for counterfactual or statistical dependencies between these variables. The arrows are placeholders for underlying causal mechanisms (Pearl, 2000) and render the variables into causes and effects. Figure 1 shows an example of a common-cause network that relates a cause feature F_C to three effect features F_{E1}, F_{E2}, and F_{E3}. The features of a category are usually coded such that the typical feature value is 1 (i.e., presence) and the atypical value is 0 (i.e., absence).

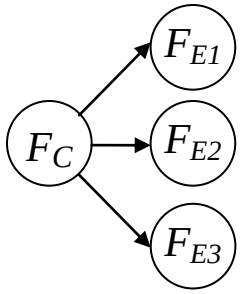

Figure 1: An example of a simple common-cause structure that connects a cause feature F_C with three effect features F_{E1}, F_{E2}, and F_{E3}. Due to the causal relations, the state of each effect feature depends counterfactually or statistically upon the state of the cause feature.

Nowadays, it's quite uncontroversial that causal knowledge is an important part of people's concepts that underlie category representation (see Rehder, 2010, for a review). But it is still in controversial debate *how* causal knowledge affects the classification of objects.

In a typical causal-based categorization task people are introduced to a target category that possesses a set of mostly three or four features. In addition, it is pointed out how these features are causally related to each other due to some causal mechanisms that hold for the category (e.g., a common-cause model as shown in Figure 1). Then, participants are presented with several potential exemplars with the category's features being either present or absent. For each of the presented exemplars, membership ratings are obtained. The enduring controversy, then, spans around the question how the instructed causal model interacts with the presence and

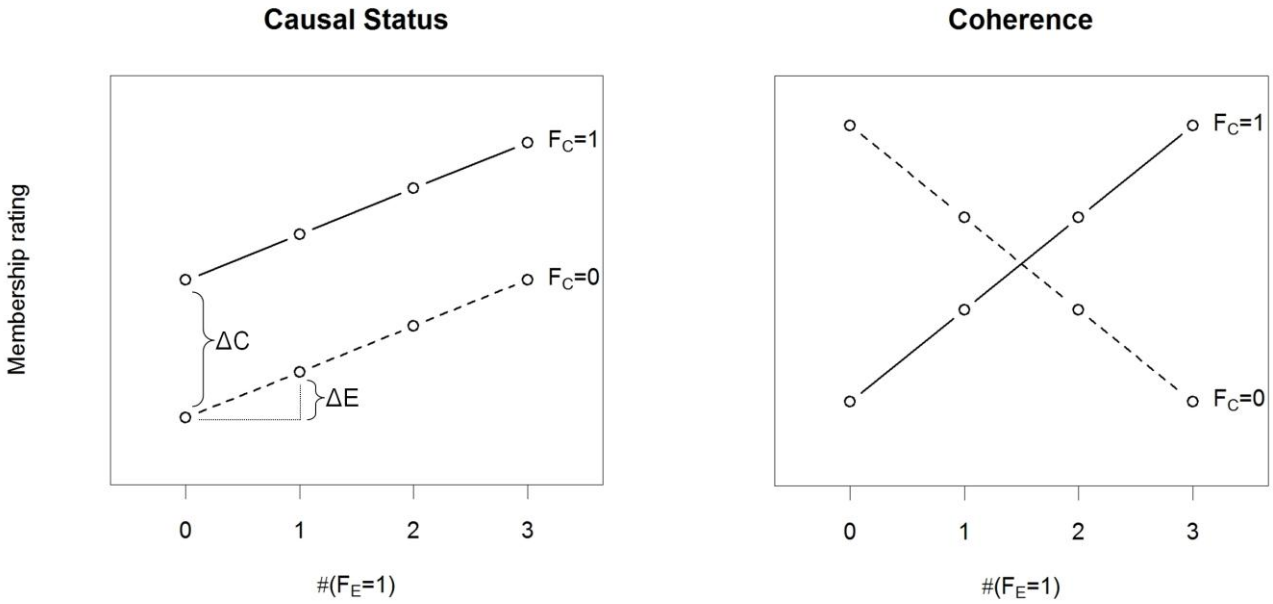

Figure 2. Predicted patterns for category membership ratings are shown according to (a) the *causal status hypothesis* and (b) the *coherence hypothesis*. The predicted ratings are computed for a category possessing four features that are connected as a common-cause model (as shown in Figure 1). The ratings depend upon the presence of the cause feature (F_C = 0 vs. F_C = 1; dashed vs. solid lines) and upon the number of effect features being present, #(F_E = 1)={0, 1, 2, 3}.

absence of features with respect to the membership ratings of the presented exemplars.

Some researchers propose that causal knowledge is an important determinant of individual feature weights in these judgments (e.g., Ahn, Kim, Lassaline, & Dennis, 2000; Marsh & Ahn, 2006). According to this account, the *causal status* of a feature matters. Others, however, emphasize the importance of feature configurations that are more or less *coherent* with the category's causal model (Rehder, 2003a, 2003b; Rehder & Hastie, 2001; Rehder & Kim, 2010). Although the effects are conceptually independent of each other (Rehder, 2010), both sides claim (and have—puzzlingly—shown empirically) that the other effect plays only a marginal, if any, role in human categorization.

In the following sections, we will describe the causal status hypothesis and the coherence hypothesis in more detail. Then, we will offer an account that—in our view—makes sense of the diverging evidence, and present an experiment that tests our claims.

Causal status effect

According to the causal status hypothesis, features that are causally more central (i.e., that have more dependents in the causal model or that appear earlier in a causal chain) have greater influence in categorization decisions when other perceptual (e.g., salience) or statistical (e.g., cue validity) properties of the material are held constant or controlled for (e.g., Ahn et al., 2000). The causal status of a feature is, therefore, an important determinant of its decision weight. For example, with respect to a common-cause model (see Figure 1), the presence vs. absence of the cause feature F_C should have more influence on the membership rating of an exemplar than the presence vs. absence of an effect feature. In Figure 2a, such an idealized causal status effect is shown: The membership ratings increase with the number of features being present; the increase, however, is higher for the cause feature (Δ*C*), than for effect features (Δ*E*).

Conceptually, the causal status effect has been linked to psychological essentialism (Ahn et al., 2000). Hence, people believe in things having essences that make them the objects they are. An essence, then, is the (unobservable) root cause for the surface features that can be observed in category members (Gelman & Wellman, 1991). Features that have a high causal status might be seen as most diagnostic for an object's essence and, therefore, category membership (Ahn et al., 2000; see also Rehder & Kim, 2010).

Coherence effect

Whereas the causal status effect is defined with respect to the weight of individual features, the coherence effect arises from the impact of feature interactions (Rehder, 2010). According to the coherence hypothesis, exemplars whose feature configurations are most coherent with the category's causal model are seen as the best members. Causal models, therefore, provide us with information about which features should go together in an exemplar. Features that are connected by a causal link should be either both present, or both absent (Rehder, 2003a; Rehder, 2010). With respect to a

common-cause model (see Figure 1), for example, membership ratings should be an increasing function of the number of effects being present when the cause feature is present, but a decreasing function in its absence. Figure 2b shows such an idealized coherence effect. Membership ratings are expected to be highest when all features are either present or absent. In this case, all causal links are preserved (i.e., such an exemplar is most coherent). The worst (i.e., most incoherent) exemplars, in contrast, are those that preserve none of the links: The cause feature is present but all effect features are absent, or the cause feature is absent but all effect features are present (three violated links in both cases).

Coherence in the proposed manner, however, faces a problem when assessed intuitively in real world cases that pop into one's mind: It does not make sense. An animal that does not have wings, does not fly, and does not build nests on trees is not a bird, although the absence of these features is perfectly coherent with the causal model of the concept "bird". Marsh and Ahn (2006) therefore suggested that the coherence effect may be not more than an experimental artifact arising from the artificial material used by Rehder and colleagues (e.g., Rehder 2003a, 2003b; Rehder & Hastie, 2001). Nevertheless, we propose otherwise.

Explanatory roles of causal models

So far, the whole debate has neglected the fact that causal models may play different roles in the representation of concepts that underlie categories. Causation in those models is implemented (or thought) in a way that causes, when present, have the power to bring about their effects, but are causally inactive when absent (Cheng, 1997; Rehder, 2003a). At first glance, this might not matter anyway: Usually, the presence of the cause goes together with the presence of its effects, as well as the absence of the cause goes together with the absence of its effects (Note, that this superficial symmetry is the basis for the coherence hypothesis). However, whereas the first fact is a direct consequence of causal mechanism, the latter is just an indirect "side effect" of it (Dowe, 2000, aptly mentioned that in this case the absence of the cause prevents the presence of its effect by omission, i.e., just by not causing it). Although this difference hasn't received much attention yet, we think it is crucial for understanding causal-based categorization.

Categorization can be seen as a kind of inference to the best explanation (Jameson & Gentner, 2008; Lombrozzo, 2009; Murphy & Medin, 1985; Rips, 1989), according to which causal models provide a system of explanatory links that tie the features of a category together. Therefore, exemplars whose configuration of features can be best explained in the light of the category's causal concept are rated as the best members. With respect to the causal analysis given above, we propose—in contrast to the coherence hypothesis—that only those feature combinations matter that are relevant with respect to the underlying causal mechanisms (e.g., when both are present, but not when both are absent), because it is the mechanism but not the regularity that has explanatory value (see Keil, 2006, for an overview).

From this point of view, we can at least differentiate two explanatory roles of causal models. First, when causal mechanisms are established in terms of typical features, the causal model serves as an explanation why the category's prototype or prototypical exemplar (i.e., all features present) is as it is. The bird example given above belongs to this type of explanation. With respect to a common-cause model (see Figure 1), we would expect a strong increase of membership ratings with more effect features being present when the cause feature is also present. In this case more and more explanatory links are served (i.e., the exemplar becomes more and more coherent with the provided explanation). But when the cause is absent, the effect features are conceptually unrelated to each other. Therefore, we would expect a much smaller increase when more and more effects are present. Since the presence vs. absence of the cause feature modulates the positive influence of the effect features, we expect membership ratings that—in the aggregate—exhibit a strong causal status effect.

Second, however, it is also possible to establish causal mechanisms with respect to atypical feature values (usually coded as absences). In this case, the causal model serves as an explanation for why a category member might deviate from the category prototype (e.g., fouling that makes an apple not looking like an apple anymore). When now presented with an exemplar that lacks all typical features, you would probably be much more willing to judge this exemplar a category member than in the bird example, because the causal model provided you with an explanation why this atypical exemplar deviates from the prototype. Thus, in case the causal model links atypical feature values, we expect a pattern that looks quite like the prediction of the coherence hypothesis (see Figure 2b). First related evidence for this proposal comes from Ahn, Novick, and Kim (2003): In their studies, participants judged persons who showed a set of abnormal characteristics (e.g., suffering from insomnia, memory deficits, and episodes of extreme anxiety) as more "normal" when provided with plausible causal relations between these abnormal characteristics compared to a condition in which no such links were provided.

To summarize, we believe that the diverging evidence found in the literature regarding the causal status and the coherence effect stems from the fact that causal models play different roles in categorization and that those studies might differ with respect to the explanatory role of the instructed causal model. In the next section we present an experiment that tests our claim.

Experiment

To test our hypothesis we adapted the material used in the experiments of Rehder (2003a; in similar versions also used in Marsh & Ahn, 2006; Rehder, 2003b; Rehder & Kim, 2008, 2010, and others). Rehder presented subjects with instructions about several artificial categories (e.g., Kehoe Ants, Mya Stars) possessing four features that were linked in a common-cause model (see Figure 1). Features were introduced without giving precise base rate information

(e.g., "Some Kehoe Ants have blood that is very high in iron sulfate. Others have blood that has normal levels of iron sulfate.", "Some Kehoe Ants have an immune system that is hyperactive. Others have a normal immune system.")[1]. Then, causal mechanisms were introduced by plausible descriptions (e.g., "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."). After participants have learned the category, they had to rate all possible exemplars (all possible combinations of features being present or absent) regarding their category membership. In his studies, Rehder found evidence for the coherence effect (but see Marsh & Ahn, 2006, for a critical discussion).

In our experiment, we used the same procedure and same material. To manipulate the explanatory role of the instructed causal model, we explicitly instructed which of the feature values were described as *typical* for the category (e.g., hyperactive immune system or normal immune system). So, between conditions, the typicality of the feature values changed but the description of causal mechanisms remained constant for the same feature values. By that, however, we manipulated the explanatory role of the causal mechanisms (i.e., whether *typical* or *atypical* values were linked by mechanisms). Furthermore, we added a replication condition that was identical to Rehder (2003a), to ensure that our procedure (and translated material) yields the same findings.

Method

Participants 96 students (62 women, mean age 22.4 years) from the University of Göttingen, Germany, participated in this experiment as part of a series of various unrelated computer-based experiments in our computer lab. Participants received either course credit or were paid €8 per hour.

Material Two categories used in Rehder (2003a) were translated into German: Kehoe Ants (a biological kind) and Mya Stars (a non-living natural kind).[2] Each category possessed four binary features. Depending on condition, each feature had a typical value (coded throughout this paper as "1") and an atypical value (coded as "0"). For example, Kehoe Ants have an immune system that was either hyperactive or working normal. Which of the two values was described as typical depended on the experimental condition (e.g., in the *typical* condition, it was stated: "Typically, Kehoe Ants have an immune system that is hyperactive. A few have a normal immune system.", in the *atypical* condition, it was stated: "Typically, Kehoe Ants have a normal immune system. A few have an immune system that is hyperactive.").

Additionally, the features were causally linked in a common-cause network. Each causal relationship was described as one feature causing another in the same way Rehder (2003a) did (see above). The description of causal mechanism was identical for all conditions.

Procedure Participants were randomly assigned to one of the two categories and to the *typical*, *atypical*, or *replication* condition. They completed the experiment individually on desktop computers. The experiment consisted of two phases, an instruction phase and a test phase.

In the instruction phase, we presented subjects with information about the category (Kehoe Ants, Mya Stars). Subjects were introduced to the four binary features and their typical values (depending on *typical* or *atypical* condition, respectively). Then, subjects were provided with information about how the features are causally connected. (As stated above, in all conditions the causal links where instructed between the same feature values. However, the typicality of these feature values and, therefore, the explanatory role of the causal model differed depending on condition.) In the *replication* condition the causal links were presented in the same way, but no information about the typicality of the values was given (as in Rehder, 2003a). The instructions were followed by a multiple choice test in which participants were required to demonstrate that they had learned all given information about the assigned category. In case of incorrect answers they had to reread the instructions and had to take the test again until they committed 0 errors.

In the test phase, subjects were presented sequentially with all 16 possible exemplars (all combinations of the four binary features) in two consecutive blocks. Order of exemplars was randomized in each block. For each exemplar, subjects were requested to give a category membership

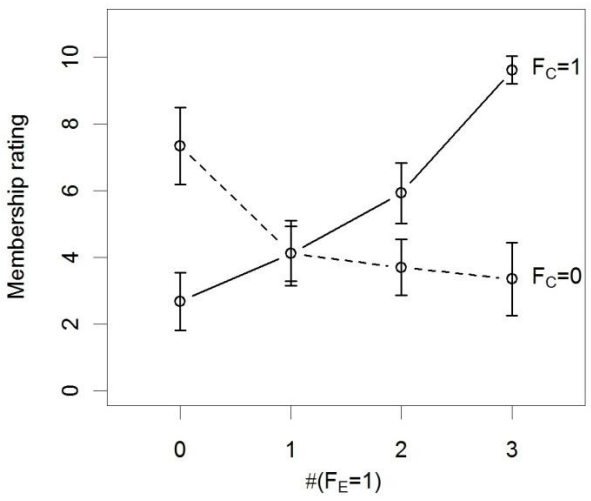

Figure 3. Results of the replication of the experiment of Rehder (2003a). Membership ratings of exemplars are shown with cause feature present (F_C=1) vs. absent (F_C=0). X-axis displays the number of effect features being present (i.e., having typical value). Error bars indicate 95% confidence intervals.

[1] Note that normal values were coded as "0" (= absence).

[2] Rehder (2003a) used six categories, but no differences for membership ratings were obtained. Therefore, we only used two randomly chosen categories.

rating on a scale from 0 (not a member at all) to 10 (definitely a member).

Design The membership ratings were aggregated for each subject across blocks and with respect to the cause feature being present (F_C = 0 vs. F_C = 1) and the number of effect features being present (#[F_E = 1] = {0, 1, 2, or 3}). This yielded a 3 (*typical*, *atypical*, *replication* condition) × 2 (category: Kehoe Ants vs. Mya Stars) x 2 (F_C = 0 vs. F_C = 1) × 4 (#[F_E = 1] = {0, 1, 2, or 3}) ANOVA design with condition and category as between-subjects factors and the presence of the cause or effect features, respectively, as within-subjects factors and average membership rating as dependent variable.

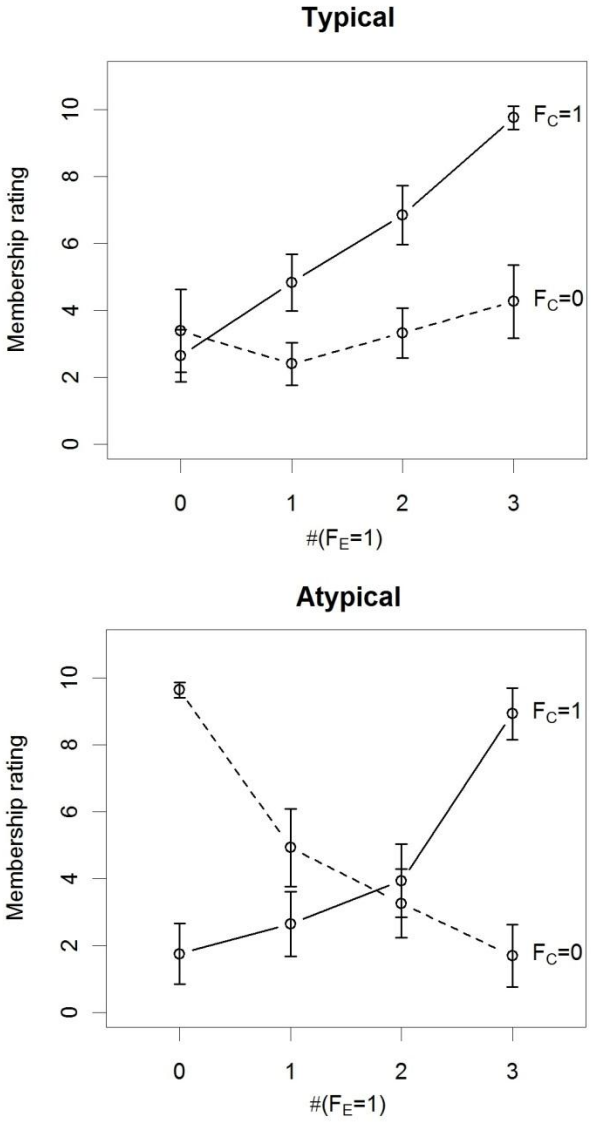

Figure 4. Average category membership ratings are shown for (a) *typical* condition and (b) *atypical* condition with cause feature present (F_C=1) vs. absent (F_C=0). X-axis displays the number of effect features being present (i.e., having typical value). Error bars indicate 95% confidence intervals.

Results

Category In line with previous studies, the category (Kehoe ants, Mya stars) revealed no significant influence on membership ratings, neither as main effect nor in any interaction (all *p*s>.1). Therefore, the factor is aggregated over in the following analyses.

Replication In Figure 3, the results of the replication condition are shown. When the cause feature was present (F_C = 1) membership ratings increased with increasing number of effects being present. In contrast, when the cause feature was absent (F_C = 0) membership ratings decreased, yielding a significant interaction ($F_{3,99}$=62.5, p<.001). This replicates the findings of Rehder (2003a) and is in line with other studies using the same material.

Explanatory Role Figure 4 displays the aggregated membership ratings for the *typical* and *atypical* condition. In both *typical* and *atypical* conditions, subjects rated exemplars with a present cause feature (F_C = 1) better members the more typical effect features were present. So, the exemplar with all features being present was rated very high (9.8 and 8.9 in *typical* and *atypical* condition, respectively) whereas the exemplar with all effects being absent was rated very low (2.6 and 1.7, respectively).

Exemplars, however, with the cause feature being absent (F_C = 0) exhibit a significant interaction between conditions ($F_{3,198}$=28.2, p<.001) (The three-way interaction was also significant, $F_{3,198}$=17.79, p<.001). In *typical* condition, ratings' increase was only marginal significant ($F_{3,102}$=2.24, p=.088). In *atypical* condition, subjects rated exemplars lower, the more effect features expressed the typical value. So, the exemplar with all effects being absent (and, therefore, all features being absent) was rated very high (9.6), whereas the exemplar with all effects being present was rated very low (1.7).

Thus, the *atypical* condition looks like the prototypical case of a coherence effect. In fact, individual influence of features (i.e., marginalized across the states of the other features) are negligible and even negative (ΔC = –0.67, ΔE = –0.14). In contrast, the *typical* condition revealed a strong causal status effect (ΔC = 2.83 vs. ΔE = 1.37), as we have predicted.

Discussion & Summary

It is widely accepted that causal knowledge is an important part of people's concepts that underlie category representations. Nevertheless, it is still quite controversial how causal knowledge affects membership ratings: Some researchers propose that causal knowledge determines the individual feature weights in categorization judgments (causal status hypothesis; see Ahn et al., 2000), whereas others, however, emphasize the role of feature combinations and whether those are coherent with the statistical regularities imposed by the category's causal model (coherence hypothesis; see e.g., Rehder, 2003a, 2003b; Rehder & Kim, 2010). We presented one possible solution to this puzzle: According to

our proposal, causal knowledge is important in categorization because it provides people with explanatory links such that they can make sense of presented exemplars. Thus, categorization is seen as inference to the best explanation (Murphy & Medin, 1985; Rips, 1989). And because the explanatory value of causal knowledge is engrained in people's beliefs about underlying mechanisms (and not statistical regularities), we derived at least two possible explanatory roles of causal models: First, a causal model can serve as explanation why a prototypical exemplar is as it is (e.g., why most birds can fly). Second, a causal model can also serve as explanation why a category member might deviate from the prototypical exemplar (e.g., why some birds cannot fly). Depending on which kind of causal model people have in mind for a given category we expect people to judge different exemplars as best and worst category members.

We presented an experiment in which we manipulated the explanatory role of the instructed causal knowledge directly, and we found huge differences in membership ratings. Interestingly, in the *typical* condition (i.e., typical feature values were linked by causal mechanisms) judgments exhibited a causal status effect. In contrast, in the *atypical* condition (i.e., atypical feature values were link by causal mechanisms) we found a strong coherence effect. Therefore, we believe that causal-status as well as coherence effects are both faces of the same coin.

Acknowledgments

This research was supported by a research grant of the Deutsche Forschungsgemeinschaft (DFG Wa 621/20).

References

Ahn, W.-K., 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.-K., Marsh, J. K., Luhmann, C. C., & Lee, K. (2002). Effect of theory-based feature correlations on typicality judgments. *Memory & Cognition*, *30*, 107-118.

Ahn, W.-K., Novick, L. R., & Kim, N. S. (2003). Understanding behavior makes it more normal. *Psychonomic Bulletin & Review*, *10*, 746-752.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149-178.

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

Dowe, P. (2000). *Physical causation*. Cambridge, UK: Cambridge University Press.

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

Jameson, J., & Gentner, D. (2008). Causal status and explanatory goodness in categorization. *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 291-296).

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, *57*, 227-254.

Lombrozo, T. (2009). Explanation and categorization: How "why?" informs "what?" *Cognition*, *110*, 248-253.

Marsh, J. K., & Ahn, W.-K. (2006). The role of causal status versus inter-feature links in feature weighting. *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 561-566).

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

Nosofsky, R. M. (1986). Attention, similarity, and the identification categorization relationship. *JEP: General*, *115*, 39-57.

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

Rehder, B. (2003a). A causal-model theory of conceptual representation and categorization. *JEP:LMC*, *29*, 1141-1159.

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

Rehder, B. (2010). Causal-based classification: A review. In B. Ross (Ed.), *The Psychology of Learning and Motivation (52)*, 39-116.

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

Rehder, B., & Kim, S. (2008). The role of coherence in causal-based categorization. *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 285-290).

Rehder, B., & Kim, S. (2010). Causal status and coherence in causal-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 1171-1206.

Rips, L. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21-59). Cambridge, UK: Cambridge University Press.

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

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

Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal knowledge and the acquisition of category structure. *JEP: General*, *124*, 181-206.