Mem Cogn (2016) 44:789–805
DOI 10.3758/s13421-016-0592-0

The diversity effect in diagnostic reasoning

Felix G. Rebitschek[1] · **Josef F. Krems**[2] · **Georg Jahn**[2]

Published online: 8 February 2016

Abstract Diagnostic reasoning draws on knowledge about effects and their potential causes. The causal-diversity effect in diagnostic reasoning normatively depends on the distribution of effects in causal structures, and thus, a psychological diversity effect could indicate whether causally structured knowledge is used in evaluating the probability of a diagnosis, if the effect were to covary with manipulations of causal structures. In four experiments, participants dealt with a quasi-medical scenario presenting symptom sets (effects) that consistently suggested a specified diagnosis (cause). The probability that the diagnosis was correct had to be rated for two opposed symptom sets that differed with regard to the symptoms' positions (proximal or diverse) in the causal structure that was initially acquired. The causal structure linking the diagnosis to the symptoms and the base rate of the diagnosis were manipulated to explore whether the diagnosis was rated as more probable for diverse than for proximal symptoms when alternative causations were more plausible (e.g., because of a lower base rate of the diagnosis in question). The results replicated the causal diversity effect in diagnostic reasoning across these conditions, but no consistent effects of structure and base rate variations were observed. Diversity effects computed in causal Bayesian networks are presented, illustrating the consequences of the structure manipulations and corroborating that a diversity effect across the different experimental manipulations is normatively justified. The observed diversity effects presumably resulted from shortcut reasoning about the possibilities of alternative causation.

Keywords Causal diversity effect · Diagnostic reasoning · Alternative causation heuristic, Suppression effect

Introduction

Humans who generate and choose diagnoses for observed symptoms use their knowledge about symptoms and potential causes, with or without reasoning about the causal development of the symptoms from the causes (e.g., Gluck & Bower, 1988). For example, knowing that hair loss is a common side effect of chemotherapy suffices to presume that a patient with hair loss underwent chemotherapy. Causal reasoning may or may not be involved in arriving at this conclusion (F. J. López, Cobos, & Caño, 2005). Research on diagnostic reasoning examines to what extent humans engage in reasoning about underlying causal processes when they aim to arrive at a diagnostic judgment.

One finding suggesting that humans may consider the processes by which observed symptoms have developed is the effect of symptoms' causal diversity when the task is to rate the probability of a proposed diagnosis (Kim & Keil, 2003). The probability of a diagnosis depends on the probability of alternative explanations, which is lower for diverse symptoms. For example, chemotherapy can cause hair loss, diarrhea, and nausea. All three are frequent consequences of chemotherapy, but pairs of these symptoms suggest chemotherapy with varying probabilities. Hair loss and diarrhea (a diverse symptom pair) suggest chemotherapy as a probable cause. Nausea and diarrhea (a proximal pair) rather suggest alternative explanations, although both are frequent effects of

✉ Felix G. Rebitschek
rebitschek@mpib-berlin.mpg.de

[1] Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

[2] Department of Psychology, Technische Universität Chemnitz, Chemnitz, Germany

chemotherapy. In the study of Kim and Keil (2003), participants learned models of causal structures and then rated proposed diagnoses higher for diverse than for nondiverse symptoms, which indicates that participants considered the causal processes. We set out to replicate this finding by Kim and Keil, and to introduce experimental manipulations that would pick up whether participants considered causal processes in this diagnostic-reasoning task.

Superior support of a hypothesis by more diverse evidence can be quantified with Bayesian causal models, as we demonstrate below. The diversity principle is a general idea in the philosophy of science (Heit, Hahn, & Feeney, 2005) and has been repeatedly confirmed in research on human thinking and reasoning. Diverse evidence produces stronger generalizations because it indicates broader categories (Homa & Vosburgh, 1976). Category-based diversity effects were shown for the evaluation of categorical arguments for induction (Osherson, Smith, Wilkie, & López, 1990), in the search for diagnostic information (Kim, Yopchick, & de Kwaadsteniet, 2008), and for the testing of arguments (A. López, 1995). Moreover, children use the diversity principle when they work through tasks of inductive reasoning (Heit & Hahn, 2001; but see Lo, Sides, Rozelle, & Osherson, 2002). Medin, Coley, Storms, and Hayes (2003) revealed that the diversity effect in inductive reasoning is influenced by retrieved knowledge.

Diversity effect by symptom-cued retrieval of alternative causes

Consider, for instance, that observed symptoms cue the retrieval of alternative causations that act as plausible counterexamples to a proposed diagnosis. Then, reasoners may rate the probability of the diagnosis in question as lower. Nondiverse, proximal symptoms that commonly appear together could cue more counterexamples than diverse symptoms (or any, instead of none), as in the chemotherapy example. An increasing number of counterexamples available in memory has been shown to increasingly suppress the acceptance of conclusions in causal conditional reasoning (De Neys, Schaeken, & D'Ydewalle, 2003), as with other types of conditionals (e.g., Byrne, 1989; Juhos, Quelhas, & Byrne, 2015). Mere memory retrieval of counterexamples, thus, would not involve considering causal processes, but would reduce the probability rating of the diagnosis in question, depending on the observed symptoms' diversity.

Provided that nondiverse symptoms cue the retrieval of counterexamples more than do diverse symptoms, enhancing retrieval by a scenario that promotes alternative causes may be one possibility for increasing the diversity effect. Imagine, for instance, that both the exemplary diverse (hair loss and diarrhea) and proximal (nausea and diarrhea) symptom pairs are observed in an emergency department instead of a cancer clinic. Then, chemotherapy is a suspected cause with a rather low base rate, and consequently it is a hardly probable explanation of diverse and proximal symptom pairs alike, because its base rate is low as compared to alternative explanations in an emergency department. Thus, normatively, a low base rate of the diagnosis in question can increase the diversity effect. After replicating the effect, one of our manipulations tested whether the psychological diversity effect in diagnostic reasoning is actually influenced by base rate information. This finding would be in line with the assumed explanation of the diversity effect by the symptom-based retrieval of counterexamples.

However, the diversity effect has not only been shown for real-life medical conditions (e.g., skin lesions and infections as a result of radiation exposure), for which more counterexamples could be cued in memory for proximal than for diverse symptoms without referring to a causal model (as for counterexamples in reasoning with causal conditionals), but also for artificial diseases and symptoms without such an asymmetry in the associated counterexamples, whose assignment to proximal and distal roles in the causal model was balanced. Memory cueing of counterexamples is thus not an exhaustive explanation of all of Kim and Keil's (2003) findings.

Diversity effect by evaluating causal processes for alternative causation

Alternatively, the diversity effect in judging the probability of diagnoses could result from evaluating causal models. The participants in the original study had learned causal chains and could have acquired knowledge representations that resembled formal descriptions of interconnected causes and effects. Alternative mental model representations for causal reasoning have been proposed (e.g., Goldvarg & Johnson-Laird, 2001); however, the materials and manipulations in the present study were aimed at producing causal models in the sense of causal network representations.

One way in which considering causal processes would produce the diversity effect can be described as *qualitative reasoning with causal models* (Sloman, 2005; Sloman & Lagnado, 2015), in which causal processes are reduced to links from causes to effects (symptoms). For example, consider the materials of Experiment 1 (see Fig. 1 and Table 1 below), which were novel in content (chemicals as root causes) but structurally equivalent to the materials with which the original finding was obtained. Two causal chains spread from the root cause, and each of the chains spreads into two effects (the symptoms). Now compare a patient who suffers from *impaired speech* and *disability of motion* with another patient suffering from *disability of motion* and *stomach ache*. The higher probability of chemical *R* as the explanation for *disability of motion* and *stomach ache* can be deduced from a simplified causal model. The causal process linking *R* and

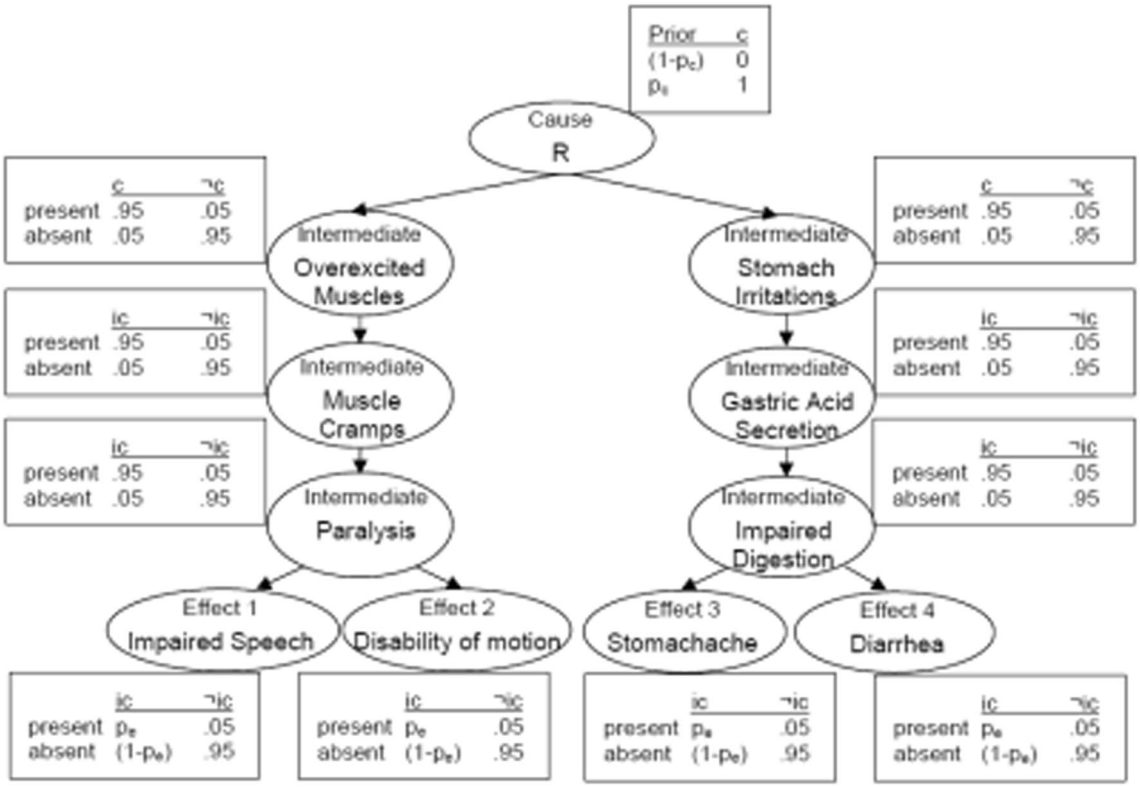

Fig. 1 Exemplary causal model with five levels (cause [c], three intermediate causes [ic], and the effects) for a single chemical in Experiment 1. Effects 1 and 2 constitute a proximal set, and Effects 3 and 4 as well. Any other pairing of effects constitutes a diverse set. The exemplary probabilities of states (present or absent) that are listed for each element were not conveyed to the participants. These probabilities were used for computing diversity effects in causal Bayes nets, as described below in the Causal Bayesian Models of a Normative Diversity Effect section

Table 1 Learning materials from Experiment 1. Participants learned about two chemicals. Each chemical was the root of two causal chains consisting of three intermediate causes (*long-chain* condition), and finally spreading into two symptoms. The causal chains in Experiments 2 and 3 and in the short-chain conditions of Experiment 4 included one intermediate cause (printed in bold) only. Items for the probability-rating task were constructed as combinations of effects. For example, a proximal symptom pair would be *disability of motion* and *impaired speech*, and a diverse pair would be *impaired speech* and *diarrhea* for the structure shown in Fig. 1

Chemical 1 (1st Level)				
Level	Intermediate Causes	Intermediate Causes	Intermediate Causes	Intermediate Causes
2nd	Overexcited muscles	Stomach irritation	Throat irritation	Impaired lymph nodes
3rd	Muscle cramps	Gastric acid secretion	Mucosa tears	Immune cell deficiency
4th	**Paralysis**	**Impaired digestion**	**Sore throat**	**Susceptibility to infection**
	Effects	Effects	Effects	Effects
5th	Disabil. of motion Impaired speech	Diarrhea Stomach ache	Bleeding throat Mucous congest.	Pneumonia Fungal disease
Chemical 2 (1st Level)				
Level	Intermediate Causes	Intermediate Causes	Intermediate	Intermediate
2nd	Bleeding	Dry eyes	Skin tingling	Allergic reaction
3rd	Blood deficiency	**Eye irritation**	**Itching**	Bronchoconstriction
4th	**Low blood pressure** Effects	Reddened eyes Effects	Scratching wounds Effects	**Asthma attack** Effects
5th	Paleness Freezing	Epiphora Impaired vision	Dermatitis Scarring	Difficult breathing Chest pain

The original materials were in German

disability of motion clearly differs from the process linking *R* and *stomach ache*, whereas both *impaired speech* and *disability of motion* are caused by *R* via a shared causal process. In this sense the *diverse* symptoms, *disability of motion* and *stomach ache*, suggest their common cause more strongly because no or hardly any alternative cause triggers both causal processes, and a synchronous elicitation of both processes by different alternative causes is unlikely, too (Kim & Keil, 2003). The root cause *R* is a parsimonious explanation (Read & Marcus-Newhall, 1993) for diverse symptoms, whereas the proximal symptom pair *impaired speech* and *disability of motion* can be parsimoniously explained by any alternative explanation for the intermediate cause *paralysis*. This asymmetry in causation is reflected in the diversity effect.

Note that considering causal processes in causal models in this way involves reasoning about alternative causations, which appears to be similar to the suppression effect by cued retrieval of counterexamples (e.g., De Neys et al., 2003), but the hypothesis's probability is not just suppressed by retrieved alternative causes. Instead, models of causal processes are evaluated, and therefore, probability judgments should be influenced by features of causal models that change both the probability of alternative causation and the causal chains.

Thus, we explored the influence of the causal structure by another experimental manipulation. As we show with exemplary Bayesian network models in a later section, base rate and causal-chain structure interact in their contributions to the normative diversity effect. The number of intermediate causes determining a causal chain's length relates to the probability of alternative causation, and thus to the diversity effect. Longer causal chains with multiple intermediate causes offer more possibilities for alternative causation. Analogously, if chains are represented as causal processes and evaluated in reasoning, a chain with more elements that invite thinking about alternative causation would lead to diverging ratings based on proximal and diverse symptoms, and thus would increase the psychological diversity effect. In contrast, the length manipulation of causal chains may leave a diversity effect unaffected that was mostly due to the retrieval-based suppression of the diagnosis in question.

Qualitative reasoning with causal models but not retrieval-based suppression explains the diversity effect for fictitious materials in Kim and Keil (2003). Kim and Keil discussed two further simple heuristics that could have produced the effect without evaluating alternative causation: Participants could have just counted whether both chains of intermediate causes are covered by the symptoms (true for diverse, false for proximal symptoms), or they could have just assigned less weight to a symptom that confirmed an already covered chain (true for proximal, false for diverse symptoms). To challenge these explanations, it is important to demonstrate that the diversity effect is sensitive to the described experimental manipulations of base rate and causal structure that change the possibilities of alternative causation but do not change the predictions for simple matching (feature-checking) or discounting heuristics.

In the presented experiments, we aimed to replicate the diversity effect in judging the probability of a proposed diagnosis. We attempted to collect evidence against the matching and discounting heuristics by demonstrating influences of base rate and chain length manipulations, and we explored the influence of considering alternative causation.

We now preview our results to set individual findings in context: Although Experiments 1, 2, and 3 indicated modulations of the diversity effect in line with causal-model theories, these were not confirmed in Experiment 4. Overall, we repeatedly observed a remarkably robust diversity effect whose unstable modulations do not strongly support theories involving causal mental models. Presumably, the observed diversity effects mainly resulted from simpler reasoning about alternative causation. They could have resulted from noncausal reasoning heuristics, as well; however, some form of causal reasoning seems likely, considering that some modulations could be observed, that participants reported representations of causal structures, and previous findings with different causal-reasoning tasks. Finally, computations of the normative diversity effect in Bayesian causal networks were conducted to illustrate the complex interplay of causal structure and base rate information and to indicate which predictions of diversity effect variations in response to the studied manipulations correspond to robust normative variations.

Experiment 1: long causal chains and high base rate

In a first step, we aimed to replicate the diversity effect in diagnostic reasoning with the original causal structure (long-chain condition; see Fig. 1), adapted for the chemical accident paradigm (Jahn & Braatz, 2014; Meder & Mayrhofer, 2013; Mehlhorn, Taatgen, Lebiere, & Krems, 2011; Rebitschek, Bocklisch, Scholz, Krems, & Jahn, 2015). Participants were told that patients showing symptoms (effects) may have been affected by a chemical (root cause) because of an accident in a chemical plant. Prior to rating the probability of the chemical as being the correct diagnosis for proximal or diverse symptoms, participants learned about the causal chains via which the chemical could cause individual symptoms. Participants never saw the complete causal structure at once (see, e.g., Fig. 1). Instead, they were presented with four causal chains, one for each of the four possible effects. This procedure prevented the probability ratings being influenced by visuospatially encoded distances between the effects. Participants could integrate the separately presented chains in order to mentally represent the underlying causal structure.

The acquired knowledge had to be reported to the experimenter, who ensured that all elements of the causal structure

were memorized and who noted whether the causal chains seemed to be integrated. Reports of integrated causal structures would directly reveal the knowledge postulated as underlying causal-model reasoning.

In each trial of the probability-rating task, symptom lists for two patients were presented. One patient showed proximal symptoms, the other showed diverse symptoms. For both patients, the probability that the chemical had caused the symptoms had to be rated. Each participant completed the learning procedure and the subsequent rating trials, first for one chemical, and then a second time for another chemical. We hypothesized that participants would rate the probability that the chemical in question had caused a patient's symptoms as higher for diverse than for proximal symptoms. Reasoning with or without causal models could produce this diversity effect. Participants whose self-reported knowledge resembled the intended causal model were expected to produce larger diversity effects than would those reporting other representations.

Method

Participants

Fifty undergraduate students (33 female, 17 male; mean age = 22.9 years, *SD* = 3.1) from the University of Greifswald participated in the experiment. Two participants were not included in the analysis because they reported professional medical experience.

Material

Learning material Table 1 shows the materials, from which causal chains of intermediate causes and symptoms were selected for each participant. The plausibility of the causal chains had been extensively pretested. Two chains from the top half of Table 1 were selected for one chemical, and two chains from the bottom half were selected for the second chemical. For the example in Fig. 1, the chains in the top left of Table 1, starting with *overexcited muscles* and *stomach irritations*, were selected.

The chemicals were referred to by single letters in the experiment. Chemical 1 was referred to as "R" or "B," and Chemical 2 was referred to as "W" or "K." In addition to the materials shown in Table 1, general symptoms (*thirst*, *tiredness*; World Health Organization, 2013) were included in the scenario. Participants were told that the general symptoms could be caused by any of the chemicals.

During learning, four causal chains linking the chemical to the single symptoms were presented on separate screens. For example, one of the learning screens for the structure in Fig. 1 showed the terms *Chemical R*, *Overexcited Muscles*, *Muscle Cramps*, *Paralysis*, and *Disability of Motion*, vertically arranged and linked by downward-pointing arrows.

Item material In each trial of the diagnostic-reasoning task, two patient vignettes were presented, vertically arranged on the right side of the screen. Each vignette consisted of a list of two or three symptoms that the respective patient suffered from. One vignette contained a diverse symptom pair, and the other contained a proximal symptom pair. For each causal structure with four effects, two proximal and four diverse symptom pairs could be constructed (see Table 2). Four of the eight possible combinations of proximal and diverse pairs were presented to a participant. In two of the four items, different general, unspecific symptoms (*thirst* or *tiredness*) were included as the third symptom in both vignettes. Next to each vignette on the left side of the screen, prompts for the probability ratings were presented: *What is the probability (on a scale from 0 to 100) that this patient had come into contact with chemical X?* with *X* being replaced by the letter of the learned chemical.

Procedure

The experiment began with the introduction of the cover story, according to which the participants as physicians had to diagnose patients who were workers in a chemical plant. It was stressed that each patient could have been affected by a chemical because of an accident, or may not have been affected by a chemical. Participants were informed that they would learn about two chemicals processed in the plant. Then, the two general symptoms that could be caused by all chemicals were introduced. Finally, the subsequent learning procedure of symptoms ("which further symptoms the chemical can cause") was announced.

In the learning procedure, a starting slide was presented that explained the structure of the learning slides to the participant and announced that the acquired knowledge had to be reported after studying. Then, four slides with the chains for the first chemical were presented. Participants studied the individual slides at their own pace. A fifth slide informed them that they could restart the presentation. Studying the four slides could be repeated as often as the participant wanted to. The presentation order of the chain slides was pseudo-randomized so that chains sharing the same intermediate causes were never presented consecutively.

After studying the chain slides, participants were invited to report the acquired knowledge to the experimenter. This studying and reporting had to be repeated until reporting was once complete. All elements contained on the slides had to be reported, but participants were free to structure and report the acquired knowledge according to their own preferences—there were no requirements regarding the structural arrangement, the characteristics of relationships (e.g.,

Table 2 Exemplary (see Fig. 1) symptom pairs. Each proximal pair could be combined with each diverse pair in a probability-rating item; four out of the eight possible combinations were selected (pseudo-randomized) and presented to a participant

Proximal Symptom Pairs			Diverse Symptom Pairs		
Impaired speech	–	Disability of motion	Impaired speech	–	Stomach ache
Stomach ache	–	Diarrhea	Impaired speech	–	Diarrhea
			Disability of motion	–	Stomach ache
			Disability of motion	–	Diarrhea

temporal or causal), and the causal links. The spontaneously expressed descriptions of what a participant had memorized from the slides were simultaneously drawn as words and arrows by the experimenter to document the verbal report. These drawings were judged as either showing the integrated causal structures underlying the material or another structural arrangement.

Diagnostic-reasoning task When learning and reporting were completed for the first chemical, instructions for the diagnostic-reasoning task were presented. Participants were informed that they would be presented with pairs of patients and that each patient would show two or three symptoms. Symptoms that were not shown were defined as truly absent. Participants were instructed to rate the probability that each patient had come into contact with the chemical.

Then, participants were presented with two patient vignettes, each containing two or three symptoms. For example, the first vignette could be *impaired speech, disability of motion, thirst*, and the second vignette could be *disability of motion, tiredness, diarrhea*. The vertical arrangement of vignettes with proximal and diverse pairs was balanced, and the vertical arrangement of the symptoms within vignettes was randomized. At the left side of the screen, next to each vignette, rating prompts were shown: *What is the probability (on a scale from 0 to 100) that this patient had come into contact with chemical X?*, with *X* replaced with the letter of the learned chemical. A sentence at the top of the screen instructed participants to consider both patients before hitting the space bar to proceed to rating. Hitting the space bar opened an input field for the rating of the first vignette next to the prompt sentence. The numerical answer was entered via the number keys on a standard keyboard. Editing with the backspace was possible. Hitting the return key opened the input field for the second vignette, and when the second rating was completed, the next item was shown.

Four items with proximal- and diverse-symptom vignettes had to be rated. Each of the two proximal-symptom pairs was presented two times with a diverse-symptom pair that was randomly drawn (with replacement) from the four possible diverse pairs (Table 2). Every second item that was presented included different unspecific symptoms (*thirst* in one vignette, *tiredness* in the other).

After four rating items, participants were again asked to report the knowledge they had acquired at the beginning. Then they worked through the learning phase and the diagnostic-reasoning task for the second chemical. The procedure was the same as for the first chemical.

The assignment of causal chains to the two chemicals (Table 1) was counterbalanced across participants, as was the order of chemicals in the experiment. For Chemical 1, two out of the four chains beginning with the intermediate causes *overexcited muscles*, *stomach irritations*, *throat irritations*, or *impaired lymph nodes* were selected. For Chemical 2, two out of the chains beginning with the intermediate causes *bleeding*, *dry eyes*, *skin tingling*, or *allergic reaction* were selected. Each chain was presented to 24 participants, each chain pair was presented to eight participants, and each combination of chain pairs was presented to two participants.

In total, probability ratings were collected for eight items (eight diverse-symptom sets and eight proximal-symptom sets). The whole experiment lasted about 40 min.

Results

The aggregated probability ratings for vignettes with diverse symptoms were higher than those for vignettes with proximal symptoms, in line with the diversity effect. The mean difference in ratings was 6.1, 95 % CI [0.2, 12.0], $d = 0.30$. The effect size was smaller than the diversity effect obtained by Kim and Keil (2003, Exp. 3) using similar ratings (derived Cohen's $d = 0.82$).

For at least one of the two chemicals, the majority of participants (85 %) reported the acquired knowledge prior to the probability-rating task as an integrated causal model with two chains linked at the chemical node and each ending in two effects. They apparently had organized the acquired knowledge in a hierarchical structure, which equaled the causal structure linking the respective chemical to its effects. Just four of the participants reported a hierarchical causal structure before one of the two trials only. Table 3 shows the mean ratings and rating differences separately for the trials before

Table 3 Descriptive statistics of all experiments: Means of participants' diverse-symptoms ratings, proximal-symptoms ratings, and mean differences between diverse- and proximal-symptoms ratings, separately for trials before which participants' report had been similar to an integrated causal model and for the remaining trials

		High-Base-Rate Conditions				Low-Base-Rate Conditions			
Long-chain conditions	Report	Exp. 1		Exp. 4				Exp. 4	
	Causal model	yes $N = 41$	no $N = 11$	yes $N = 40$	no $N = 8$			yes $N = 34$	no $N = 14$
		M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)			*M* (*SD*)	*M* (*SD*)
	Diverse	71.0 (21.1)	68.7 (17.8)	68.9 (17.1)	62.1 (10.5)			69.9 (6.4)	68.8 (20.0)
	Proximal	63.4 (21.2)	70.6 (20.9)	61.2 (17.7)	57.9 (14.8)			63.5 (20.6)	56.6 (22.9)
	Difference	7.5 (21.5)	–1.9 (23.3)	7.7 (18.7)	4.3 (11.0)			6.4 (19.1)	12.2 (15.7)
Short-chain conditions	Report	Exp. 2		Exp. 4		Exp. 3		Exp. 4	
	Causal model	yes $N = 43$	no $N = 11$	yes $N = 40$	no $N = 8$	yes $N = 42$	no $N = 9$	yes $N = 40$	no $N = 8$
		M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
	Diverse	66.5 (17.1)	66.8 (16.3)	73.9 (17.1)	70.1 (9.7)	69.7 (16.5)	61.0 (21.6)	67.4 (19.8)	70.4 (24.9)
	Proximal	61.1 (18.2)	70.3 (17.9)	65.9 (15.7)	70.9 (12.6)	61.7 (17.7)	52.6 (22.8)	60.5 (23.7)	61.2 (15.9)
	Difference	5.4 (16.7)	–3.5 (22.0)	8.0 (14.6)	–0.8 (4.6)	8.0 (21.3)	8.4 (15.3)	6.9 (22.8)	9.3 (24.8)

which the learned knowledge had been reported as an integrated causal model, as well as for the remaining trials. As is shown in the top left pair of columns in Table 3, the diversity effect was restricted to the trials before which participants had reported integrated causal models (respective column with the head *yes*), with a mean difference between the ratings of proximal and diverse symptoms of 7.5, 95 % CI [0.7, 14.3], d = 0.35. The remaining trials (column with the head *no*) did not show a diversity effect (d = –0.08).

The diversity effect did not seem to be restricted to subsets of the material. A two-way analysis of variance (ANOVA) including the factors Symptom Diversity and Causal Chain of the proximal symptom pair (Table 1) did not indicate a diversity-by-material interaction effect on ratings, $F(7, 184) = 1.22$, $p = .29$; the main effect of diversity was confirmed, $F(1, 184) = 12.36$, $p = .001$, $\eta_p^2 = .06$, without showing a main effect of the material on the ratings, $F(7, 184) = 1.86$, $p = .08$.

Finally, notice that ratings based on three symptoms (including an unspecific symptom) were higher than those based on two symptoms (d = 0.37). This may indicate that unspecific symptoms decreased the subjective probability of alternative causations that were not mentioned in the scenario. Nonetheless, the diversity effect based on three symptoms (d = 0.26) and the effect based on two symptoms (d = 0.31) were of similar sizes.

Discussion

The diversity effect was confirmed, and moreover, it was restricted to trials before which participants had integrated the causal chains during learning. In rating the probability of a diagnosis for observed symptoms, these participants seemed to engage in reasoning about alternative causation that may have involved a causal model. The proposed diagnosis (contact with the chemical in question) was a parsimonious explanation of diverse symptoms. For the proximal symptoms, alternative causation of the learned intermediate causes in the respective chain could have been imagined, which then explained the proximal symptoms parsimoniously (Kim & Keil, 2003). Although those participants who showed the diversity effect reported causal models, they could have retrieved alternative causes of the respective symptom constellations without considering causal processes. If alternative explanations were easier to retrieve for proximal symptoms, this could have suppressed the rating of the proposed diagnosis more for proximal than for diverse symptom pairs (cf. De Neys et al., 2003).

Manipulations that affect reasoning about alternative causation with causal models but do not affect reasoning about alternative causation based just on symptom-induced memory retrieval can differentiate between those variants of causal reasoning. Thus, in Experiment 2, the causal chains were shortened, and only the intermediate causes that are printed in boldface in Table 1 were retained on the learning slides (*short-chain* condition). A reduced structure with less probabilistic elements implies more deterministic cause–effect relationships. Thus, the root cause is promoted, and the possibilities of alternative causation are reduced. The probability of the root cause is increased primarily for proximal symptoms that, with longer chains, could have more easily prompted the retrieval of alternatives than

diverse symptoms. As a consequence, the probabilities of the root cause given diverse and given proximal symptoms would become more similar with shorter chains, and the diversity effect should be reduced.

Experiment 2: Short causal chains and high base rate

The procedure in Experiment 2 was the same as that in the previous experiment, except for the reduction of the intermediate causes in the causal chains from three to one (*short-chain* condition), for which alternative explanations could be imagined. Therefore, the same vignettes as in Experiment 1 could be presented for probability ratings, and again we hypothesized that participants would rate the probability of contact with the respective chemical as being higher for diverse than for proximal symptoms. Shortening the chains, however, would increase the likelihood that a present symptom was brought about by the root cause, which, first, should be reflected in generally increased ratings of the root cause. Second, if participants applied reasoning about alternative causation (cf. Kim & Keil, 2003) in a way that was sensitive to the causal-structure manipulation, the diversity effect should be reduced from Experiment 1 (long chains) to Experiment 2 (short chains). With shortened causal chains, alternative explanations for proximal symptoms were harder to imagine than with multiple intermediate causes in long chains. Thus, the ratings for diverse and proximal symptoms should converge.

Method

Participants

Forty-nine undergraduate students (39 female, 10 male; mean age = 22.6 years, *SD* = 2.8) from Chemnitz University of Technology participated in the experiment. One participant was not included in the analysis because she reported professional medical experience.

Material

Learning material The causal chains from the top and bottom of Table 1 were assigned to Chemicals 1 and 2, respectively. However, the causal chains contained only one intermediate cause (printed in boldface in Table 1). The plausibility of the reduced causal chains had been pretested.

Experimental material The lists of symptoms and their assignments to the proximal and diverse vignettes, the vignettes' spatial arrangements, and the symptoms' spatial arrangements were the same as in Experiment 1. The assignment of causal chains to chemicals and the order of chemicals in the experiment were again counterbalanced across participants.

Procedure

The procedure was the same as in Experiment 1. Experiment 2 lasted about 30 min.

Results

The diversity effect in probability ratings was not statistically significant and tended to be reduced as compared with Experiment 1 (Fig. 2, left diagram). For the whole group of participants, the mean difference between the aggregated ratings of diverse and proximal symptoms was 3.4, 95 % CI [–1.7, 8.6], $d = 0.20$.

The proportion of participants whose reports of the acquired knowledge resembled a causal model for at least one of the two chemicals was 90 %. For the respective rating trials, the mean difference was slightly larger, $M = 5.4$, 95 % CI [0.3, 10.5], $d = 0.32$, and again no diversity effect was observed for the remaining trials, $d = –0.16$ (bottom left pair of columns in Table 3).

Symptom diversity acted similarly across the material assignments, as we confirmed in a two-way ANOVA (symptom diversity, proximal chain), checking for a diversity-by-material interaction effect on the ratings, $F(7, 184) = 0.75$, $p = .63$; the main effect of diversity was confirmed but was reduced as compared to Experiment 1, $F(1, 184) = 5.72$, $p = .018$, $\eta_p^2 = .03$, again without showing a main effect of material on the ratings, $F(7, 184) = 1.86$, $p = .08$.

As in Experiment 1, the ratings based on three symptoms (including an unspecific symptom) were higher than those based on two symptoms ($d = 0.31$), which may suggest that the subjective probability of unknown alternative causations was decreased by unspecific symptoms. Nonetheless, the diversity effect based on three symptoms ($d = 0.22$) and the effect based on two symptoms ($d = 0.16$) were of similar sizes.

Discussion

In Experiment 2, the diversity effect was apparently reduced, if it was present at all. In the context of the diversity effect in Experiment 1, it seems that causal structure information was considered in the two experiments. A represented causal structure might be necessary for causal reasoning with causal models (Meder, Hagmayer, & Waldmann, 2008, 2009). More reliable evidence for considering causal structures could be provided if such a reduction in the diversity effect was revealed by a structure manipulation within a single experiment.

Note that heuristic symptom processing, without considering alternative causation such as counting of

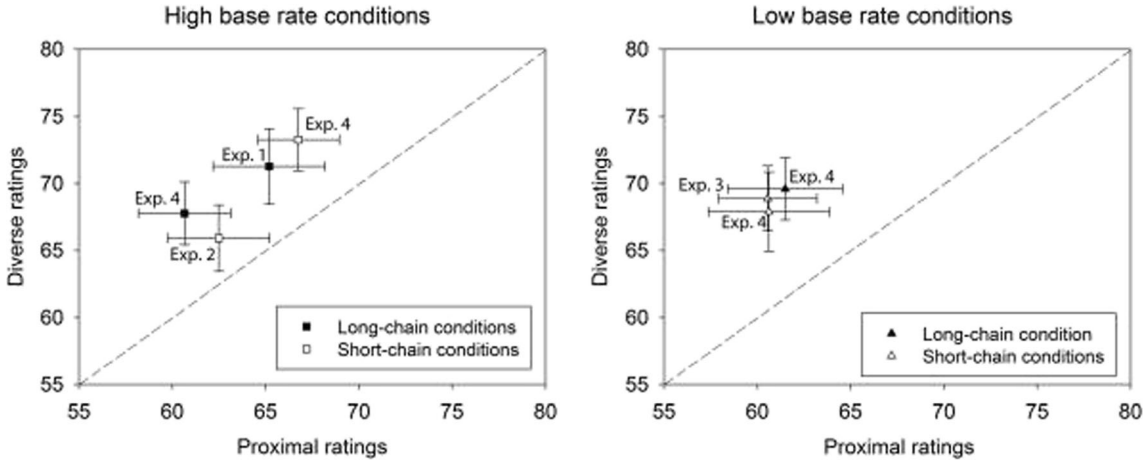

Fig. 2 Mean ratings across participants in the respective conditions of Experiments 1, 2, 3, and 4. High-base-rate conditions (left) and low-base-rate conditions (right) are separated. Error bars denote the standard errors

matched causal chains or discounted weighting of symptoms that regularly appear together (Kim & Keil, 2003), fails to explain the apparently absent diversity effect in Experiment 2, given the diversity effect with longer chains in Experiment 1. This is because such shortcut strategies do not predict effects of the causal structure manipulation. A confirmation of this cross-experimental observation could thus rule out these explanations.

Confirming the effects of causal structure in a single experiment (which we attempted in Exp. 4) is of additional importance, because the shorter chains in Experiment 2 did not result in higher probability ratings overall (Fig. 2, left diagram, and Table 3). The large interindividual variation in ratings had been expected because participants received only qualitative descriptions of the scenario and causal links. Individual participants choose ratings in a certain range based on their individual scenario interpretation, and thus the differences between ratings, but not the absolute ratings, can be reliably interpreted. Yet an increase in probability ratings for shorter chains might surface if the length of chains were varied within subjects (see Exp. 4).

Note, that we found no hint that the shortened chains were less integrated during learning, which could have happened because the small number of terms to be memorized might have reduced the need for organization. The small number of terms was reflected in the reduced number of repetitions while studying the material during learning: Participants needed 1.1 repetitions after first studying the slides in Experiment 2, but 2.4 repetitions in Experiment 1, $t(71) = 3.53$, $p = .001$. Nonetheless, the proportions of participants reporting the acquired knowledge as a causal model were stable.

Although Experiment 2 indicated that the diversity effect may be reduced by shortening the chains, the diversity effect could also have been absent because reasoning changed qualitatively; participants may not have considered alternative causation at all with the shorter chains in Experiment 2. We therefore were interested to see whether a diversity effect could be shown with shorter chains. In addition, showing that long chains are not required could reduce the learning demands in future experiments studying the diversity effect in diagnostic reasoning. If the diversity effect can be ascribed to reasoning about alternative causation, either as cued retrieval of counterexamples of causes or as imagination of alternative causes for the intermediate causes, the effect should reappear with the shortened causal structure if alternative causation becomes more likely. Hence, in Experiment 3 we examined whether the diversity effect would surface again if the base rate of the root cause were lowered, in order to generally increase the probability of alternative causation.

Experiment 3: Short causal chains and low base rate

Experiment 3 was a copy of Experiment 2 (short-chain condition), except for the cover story, which was changed to suggest a lower base rate of the root cause (chemical contact). Participants were no longer told that the patients showing diverse or proximal symptoms were workers from a chemical plant, but instead that the patients were inhabitants of a town nearby the chemical plant who were visiting their family physician.

A decreased base rate of the chemical would invite alternative causation. Thus, the ratings of the chemical cause should be generally decreased as compared with Experiment 2. Furthermore, if participants engaged in causal reasoning by considering alternative causation, this manipulation should increase the diversity effect as compared to Experiment 2. The presence of proximal symptoms would be more likely to prompt the retrieval of counterexamples that would suppress the rating of the chemical cause (De Neys et al., 2003), and the proximal symptoms' causal development could be more

easily be attributed to a single alternative cause. In contrast, diverse symptoms would still require causations of two distinct intermediate causes, and therefore would suggest the chemical as a probable diagnosis relative to alternatives.

Method

Participants

Forty-nine undergraduate students (38 female, 11 male; mean age = 23.2 years, *SD* = 3.1) from Chemnitz University of Technology participated in the experiment. One participant was not included in the analysis because he reported professional medical experience.

Material

The materials were the same as in Experiment 2. Only the intermediate causes that are printed in boldface in Table 1 were included on the learning slides for the causal chains.

Procedure

The procedure was the same as in the preceding experiments, except for the cover story, which now suggested a lower base rate of the chemical. Whereas the patients had been described as workers at the chemical plant in the preceding experiments, in Experiment 3 the patients were described as inhabitants of the town nearby the chemical plant, who were visiting a family physician and may have been affected by the chemical, but may also have suffered from all kinds of diseases. By this change, the presumed base rate of a chemical as the cause of symptoms should be decreased as compared with potential alternative causes of the patients' symptoms. The experiment lasted about 30 min.

Results

The ratings of patients' vignettes containing diverse symptoms were clearly higher than those of vignettes containing proximal symptoms (Fig. 2, right diagram). The mean difference in ratings was larger than in Experiment 2 and was statistically significant, $M = 8.3$, 95 % CI [2.5, 14.2], $d = 0.42$.

The diversity effect this time was present in trials, before which participants' reported knowledge resembled causal models (88 % of participants reported integrated causal models for at least one of the two chemicals), $M = 8.0$, 95 % CI [1.3, 14.6], $d = 0.37$, and an effect seemed to be shown in the remaining trials, as well, $d = 0.55$ (see Table 3).

The respective materials of the causal chains may have influenced the rating difference in the present experiment, as indicated by a two-way interaction of the factors Diversity and Causal Chain of the Proximal Symptoms, $F(7, 184) = 2.06$, $p = .050$, $\eta_p^2 = .07$; the main effect of diversity was confirmed, $F(1, 184) = 27.27$, $p < .001$, $\eta_p^2 = .13$, without showing a main effect of material on the ratings, $F(7, 184) = 0.48$, $p = .85$.

Ratings based on three symptoms (including an unspecific symptom) were again higher than those based on two symptoms ($d = 0.54$). Unknown alternative causations were presumably rejected when the additional, unspecific symptoms were present. In the present experiment, the diversity effect based on three symptoms ($d = 0.31$) was smaller than the effect based on two symptoms ($d = 0.48$).

Note that lowering the base rate was expected to lower the probability ratings for causation by the chemical overall, as compared to Experiment 2 with the same chain length; however, no such consistent decrease of probability ratings was observed. As in the preceding experiments, the large interindividual variation in ratings due to the qualitative descriptions of the scenario and causal links may have concealed this decrease. It might surface if the base rate were varied within a single experiment (see Exp. 4).

Discussion

By decreasing the base rate of the root cause in Experiment 3, we obtained a reliable diversity effect even with just one level of intermediate causes in the causal chains. This finding opens the possibility for studying the diversity effect more economically in the future, because in principle it can be obtained with shortened chains, and thus decreased learning demands.

In line with reasoning about alternative causation, the apparently larger difference between the diverse-symptoms and proximal-symptoms ratings in Experiment 3 than in Experiment 2 suggests that symptom diversity was more important if the presence of the root cause was generally less probable. Like the suggested reduction of the diversity effect by shortening chains in the case of a high base rate of the root case (Exp. 2 vs. Exp. 1), the suggested increase of the diversity effect by lowering the base rate (Exp. 3 vs. Exp. 2) awaited confirmation by comparisons performed within a single experiment (Exp. 4).

Contrary to the preceding experiments, in Experiment 3 the diversity effect was present regardless of whether or not participants' reports of memorized information were similar to an integrated causal model. This puts in question the suggestion from the results in Experiments 1 and 2 that an accessible causal model representation is required for the diversity effect to occur, and it weakens the consistency with causal-model theories.

Up to this point, we have indicated the variability of the causal-diversity effects across different experiments that varied the chain length in causal structures and the base rate instructions. The effect was clearly present with long chains and a high base rate (Exp. 1) and with short chains and a low base rate

(Exp. 3). This pattern so far is in line with reasoning about alternative causation, because long chains with multiple intermediate elements and a low base rate increase the probability of alternative causation and should increase the difference between ratings for diverse and for proximal symptoms, and consequently the diversity effect. In contrast, short chains and a high base rate (Exp. 2) should decrease the diversity effect, and the results so far suggest that this may be the case. In order to test whether the diversity effect indeed changes systematically, depending on chain lengths and base rates, we conducted a further experiment at a single laboratory with random assignment of participants from the same population to two base rate conditions (high vs. low) and with a within-subjects manipulation of causal structure (long vs. short causal chains).

Experiment 4: Chain length and base rate fully crossed

Experiment 4 was similar to the preceding experiments. It included a within-subjects manipulation of chain length (long vs. short, as in Exp. 1 vs. Exps. 2 and 3) and varied the cover story between subjects, creating two base rate conditions (high vs. low). Whereas participants in the high-base-rate condition were told that the patients showing diverse or proximal symptoms were workers from the chemical plant (as in Exps. 1 and 2), participants assigned to the low-base-rate condition were told that the patients were inhabitants of a town nearby the chemical plant who were visiting their family physician (as in Exp. 3).

The same vignettes as in the preceding experiments were presented for probability ratings. Again, we hypothesized that participants would rate the probability of contact with the respective chemicals as higher for diverse than for proximal symptoms. A high base rate and short chains should increase the probability of causation by the chemical. Hence, the probability ratings for both diverse and proximal symptoms should be higher with a high base rate and short chains. The probability of causation by the chemical should increase more for proximal than for diverse symptoms, because the probability of alternative causation is higher for proximal symptoms. Hence, with a high base rate and short chains (as in Exp. 2), the ratings for proximal and diverse symptoms should converge, and the diversity effect should decrease.

Method

Participants

Ninety-nine undergraduate students (77 female, 22 male; mean age = 23.7 years, *SD* = 5.1) from the University of Greifswald participated in the experiment. Three participants were not included in the analysis because they reported professional medical experience.

Material

The materials were the same as in Experiments 1, 2, and 3. In the short-chain conditions, only the intermediate causes that are printed in boldface in Table 1 were included on the learning slides for the causal chains.

Procedure

The procedure was similar to those in the preceding experiments, with a cover story that depended on the base rate condition. To participants in the high-base-rate group, the patients were described as workers from the chemical plant; but to the low-base-rate group, the patients were described as inhabitants of the town nearby the chemical plant, who were visiting a family physician and may have been affected by the chemical, but may also have suffered from all kinds of diseases. Each participant worked through one chemical with a hierarchical structure consisting of five levels (long-chain condition) and through one chemical with a structure consisting of three levels (short-chain condition). The order of these causal-chain conditions was counterbalanced. For each chemical, participants' reports were scored as suggesting or not suggesting causal models. The experiment lasted about 30 min.

Results

The ratings of patients' vignettes containing diverse symptoms were higher than those of vignettes containing proximal symptoms across all of the experimental conditions (Fig. 3).

A 2 × 2 × 2 ANOVA including the within-subjects factors Symptom Diversity (diverse vs. proximal symptom constellations) and Chain Length (long vs. short) and the between-subjects factor Base Rate (high vs. low) confirmed the main effect of diversity, $F(1, 94) = 18.07$, $p < .001$, $\eta_p^2 = .16$. In each of the four conditions combining the chain length manipulations and base rate manipulations, the ratings of patients' vignettes containing diverse symptoms were clearly higher than those of vignettes containing proximal symptoms. We observed no clear main effect of chain length, $F(1, 94) = 2.93$, $p = .090$, $\eta_p^2 = .03$, nor a main effect of base rate ($F = 0.57$). Instead, there was a two-way interaction effect of chain length and base rate on the ratings, $F(1, 94) = 7.24$, $p = .008$, $\eta_p^2 = .07$. Only the high-base-rate group proved sensitive to the chain length manipulation (Figs. 2 and 3) and gave higher probability ratings for both proximal and diverse symptoms when the chain length was short, $t(47) = 3.20$, $p = .002$, $d = 0.46$. The three-way interaction effect ($F < 1$) was not statistically significant, and the same applied to the remaining two-

way interactions of symptom diversity and base rate ($F < 1$) and of symptom diversity and causal structure ($F < 1$). The absence of any interaction effect involving diversity indicates that the diversity effect was not modified by the chain length and base rate manipulations.

Reported knowledge was judged separately for the long- and short-chain conditions. As is depicted by the filled and empty dots in Fig. 3, the diversity effect in Experiment 4 did not depend on participants' knowledge reports suggesting integrated causal models. An analysis of the long-chain conditions indicated diversity effects for both causal-model-like reports (ds = 0.41 and 0.35 for the high- and low-base-rate conditions, respectively) and other reports (d = 0.79 for the low-base-rate condition). A similar analysis of the short-chain conditions also indicated diversity effects for both causal-model-like reports (ds = 0.55 and 0.31 for the high- and low-base-rate conditions, respectively) and other reports (d = 0.39 for the low-base-rate condition).

Because participants did not forget what they had learned about the chemical of the first trial when working through the second trial based on another chemical, transfer effects cannot be excluded. Thus, the trials were analyzed separately: The two respective 2 × 2 × 2 ANOVAs neither revealed main effects for causal structure [$F(1, 92) = 2.27$, $p = .14$, for the first, and $F < 1$ for the second chemical] nor for base rate [$F < 1$ for the first, and $F(1, 92) = 1.17$, $p = .28$, for the second chemical]. The diversity effect was shown for both trials (η_p^2s = .15 and .13 for the first and the second chemicals, respectively). Moreover, neither trial-specific interactions between symptom diversity and structure [$F(1, 92) = 2.76$, $p = .10$, for the first, and $F(1, 92) = 3.41$, $p = .07$, for the second chemical] and diversity and base rate (Fs < 1 for both chemicals) were indicated, nor were the three-way interactions (Fs < 1 for both chemicals). Thus, potential order effects were not indicated.

However, it cannot be ruled out that the respective materials of the causal chains may have influenced the rating difference in the present experiment. This was indicated by a

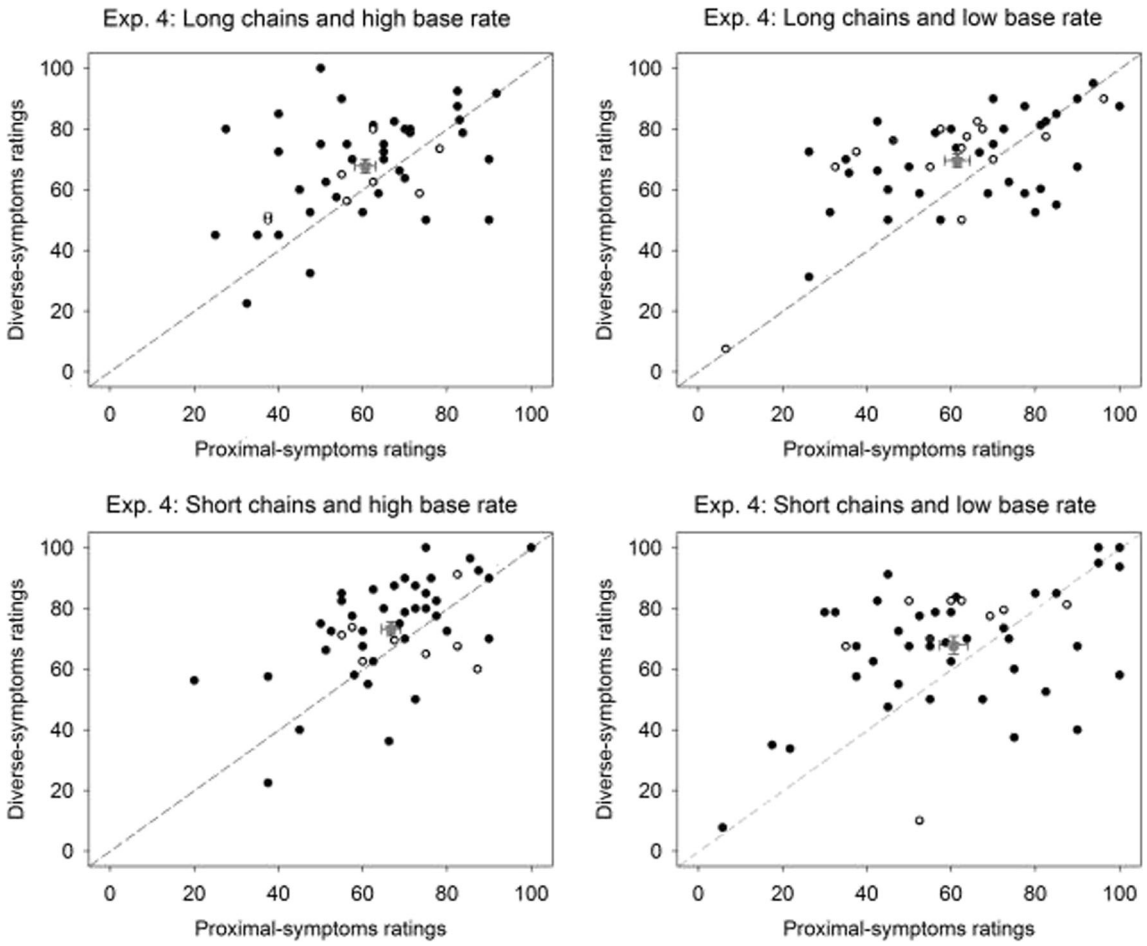

Fig. 3 Scatterplots showing mean ratings for proximal- and diverse-symptom pairs by individual participants in Experiment 4. Filled dots show participants who reported their learned knowledge similar to a causal model. Empty dots show the remaining participants, whose reports did not suggest an integrated causal mental model. Additionally, the mean ratings averaged across participants are marked as gray dots with error bars. Error bars denote the standard errors of these means

two-way interaction of the factors Diversity and Causal Chain of the Proximal Symptoms, $F(7, 376) = 2.17$, $p = .036$, $\eta_p^2 = .04$; the main effect of diversity was confirmed, $F(1, 376) = 47.30$, $p < .001$, $\eta_p^2 = .11$, without showing a main effect of material on the ratings in general, $F(7, 376) = 1.50$, $p = .16$. Finally, ratings based on three symptoms (including an unspecific symptom) were again higher than those based on two symptoms ($d = 0.62$). The additional unspecific symptoms may have helped participants reject potential alternative causations. This consideration of the unspecific symptoms is not reflected in the diversity effect. Not a single interaction of symptom diversity with the number of symptoms was statistically significant; respective 2 × 2 ANOVAs revealed no interactions for conditions with low base rate and long chains ($F = 1.18$), with high base rate and long chains ($F = 0.39$), with low base rate and short chains ($F = 0.48$), or with high base rate and short chains ($F = 1.46$).

Discussion

A reliable diversity effect was shown across conditions with short as well as with long chains and with low-base-rate as well as with high-base-rate instructions. Reasoning about the relative probabilities of diagnoses was only partially influenced by the experimental manipulations: The base rate of the root cause and the chain length to the root cause did not affect ratings systematically. Only the high-base-rate group gave higher ratings at the shorter chain length. A variation of the size of the diversity effect—for example, according to the predictions of reasoning about alternative causation with a causal model (Kim & Keil, 2003)—could not be revealed. Overall, the results did not confirm the malleability of the diversity effect by chain length and base rate manipulations that had been suggested by the previous experiments.

Furthermore, as in Experiment 3, the diversity effect was present regardless of whether or not participants' reports were similar to an integrated causal model. Thus, it can be questioned that an integrated causal structure representation is a precondition of the diversity effect.

Participants' ratings were influenced by diversity, but either the probability of alternative causation had no further differentiated effect on ratings, or such an effect was occluded by interindividual variability in interpreting the qualitatively presented information and by variability in the materials. Participants received sparse qualitative descriptions and had a wide margin for subjective interpretation of, for example, the base rates or the strengths of causal links. To illustrate that normatively, manipulations of chain length, base rates, and causal strengths interact in a complex way, below we provide exemplary computations of the normative diversity effect in Bayesian causal networks.

Causal Bayesian models of a normative diversity effect

Bayesian inference in causal networks extends the classical computation of Bayesian posterior probabilities that a certain hypothetical cause is the true cause (given the evidence and prior probabilities), with the causal processes generating evidence (Meder, Mayrhofer, & Waldmann, 2014). Bayesian networks, as directed graphical probabilistic models (Pearl, 2000), consist of edges (which are the directed causal influences) and nodes (variables reflecting the probabilistic dependencies). The conditional probability distribution for each variable depends on a set of parameters and on the respective higher-level cause. A full joint probability distribution over the causal structure is the product of all variables' conditional probability distributions. Hence, manipulating the conditional probability distributions changes the inferred posteriors in line with Bayes's theorem. The following illustration is limited to exemplary point estimates of posterior probabilities.

The models for computing normative diversity effects were constructed in analogy to the causal structures underlying the presented experiments. Two models contained two chains with three intermediate causes (as in Fig. 1; Exp. 1 and the long-chain conditions in Exp. 4); two further models contained two chains with a single intermediate cause (Exps. 2 and 3, and the short-chain conditions in Exp. 4). The node of the root cause (c) had two possible states: c (root cause present) or not c (root cause absent). Likewise, the nodes of intermediate causes (ic) and effects (Effects 1, 2, 3, and 4) were either present or absent. In all models, the effects were mutually independent but not mutually exclusive.

Furthermore, the probabilities of the different states were defined for each model. Random presence of any element but the root cause was fixed at a rate of .05. The base rate of the root cause's presence was set to $p_c = .30$ in two models, to reflect a rather high base rate of the diagnosis in question (*three ic levels and high* p_c and *one ic level and high* p_c). In two other models, the root cause's prior was fixed to $p_c = .10$, to reflect a decreased base rate (*three ic levels and low* p_c and *one ic level and low* p_c).

For manipulating causal strength, the probability that a single effect was caused by the root cause was varied between .20 and .80 (in steps of .10) for each of the models. This probability is the product of the causal strengths along the respective causal chains. The causal strength between the root cause and the first intermediate cause and the causal strength between intermediate causes was fixed at .95 (see Fig. 1). Thus, the causal strength between the last intermediate cause and the effect had to be chosen differently for short and long causal chains in order to arrive at the intended

probability.[1] In analogy to the experiments, in which the instruction "*x* can cause *y*" neither suggested a very loose nor a deterministic relationship, causal strengths beyond this range were not considered. In addition, unspecific effects (general symptoms) were defined as nondiagnostic: The probability of their presence given the root cause was the same as their probability given the alternatives. The models were built and run in a software tool for building causal models as Bayesian networks (HUGIN; for a list of academic and commercial systems, see Murphy, 2013).

For all four models, conditional posterior probabilities of the root cause's presence were inferred from diverse and proximal symptoms across the range of considered causal strengths. The diversity effect was calculated as the difference between the posterior probabilities for diverse and proximal symptoms.

Figure 4 shows that a normative diversity effect is present across the whole range of considered causal strengths for all four models. Second, the diversity effect is positively related to causal strengths up to .70, for all four models. Third, the influence of the causal structure obviously depends on the base rate of the root cause. In the case of a high base rate, the diversity effect is clearly stronger with three ic levels (filled squares) than with one ic level (empty squares). In the case of a low base rate, however, the diversity effect is stronger with one ic level (empty triangles) than with three ic levels (filled triangles), if the causal strength is about .40 or more. Hence, the interplay of structure and base rate information on posteriors and on the normative diversity effect is quite complex, even if only a selected range of parameters is considered.

Although we do not intend to interpret quantitative similarities between the posterior probabilities inferred in a Bayesian network and human probability ratings, we note that the normative diversity effect is positive across the explored parameter range (Fig. 4), as it was in all but one experiment. The explored range of parameters was limited, in that only two different base rate conditions were implemented, the causal strength with which the chemical cause was related to the individual effects was held constant across the effects, and causal strength was restricted to a range from .20 to .80. Exploring causal strengths between the cause and its effects beyond .95 revealed that the normative diversity effect can be inverted. Indeed, the effect's variation in the parameter space supports concerns about a rigid preference for diverse evidence from a Bayesian point of view (Lo et al., 2002).

[1] For example, if the intended probability was .20, the causal strength between the last intermediate cause and the effect (p_e in Fig. 1) was set to .21 for short chains (.95 × .21 = .20) and to .23 for long chains (.95 × .95 × .95 × .23 = .20). Thus, the probability of each effect's presence given the direct intermediate (p_e | ic) was varied between .21 and .84 for short chains and between .23 and .93 for long chains, to arrive at probabilities for causation by the root cause in the range between .20 and .80.

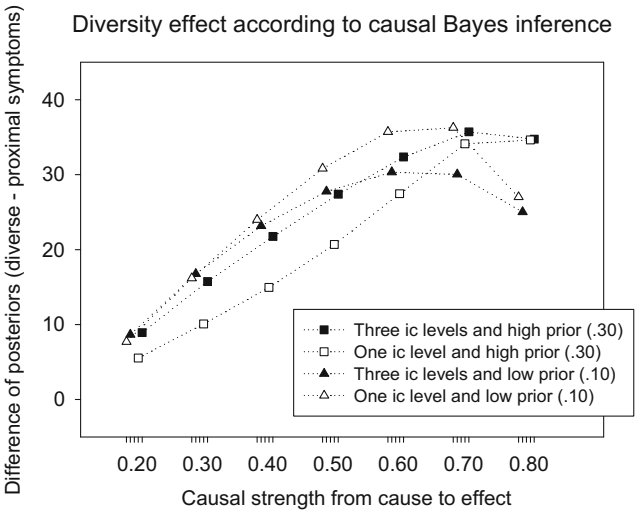

Fig. 4 Diversity effects for four exemplary causal Bayesian models, depicted as differences in posterior probabilities as percentages, computed for diverse and proximal symptoms as a function of the causal strength between the root cause and a single symptom. Diversity effects for the *three ic levels and high* p_c model (corresponding to the structure and cover story in Exp. 1 and in one condition of Exp. 4) are plotted with filled squares. Diversity effects for the *one ic level and high* p_c model (Exp. 2 and one condition of Exp. 4) are plotted with empty squares; effects for the *one ic level and low* p_c model (Exp. 3 and one condition of Exp. 4), with empty triangles; and effects for the *three ic levels and low* p_c model (one condition of Exp. 4), with filled triangles

General discussion

Diverse symptoms usually suggest fewer plausible explanations, and consequently increase the certainty in a diagnosis that explains them parsimoniously. In three out of four experiments, participants indeed rated the probability of a diagnosis as higher for diverse symptoms than for proximal symptoms when both cases were presented side by side. The first experiment conceptually replicated a diversity effect reported previously (Kim & Keil, 2003). In Experiment 2 the causal structure was modified, and in Experiment 3 the base rate of the diagnosis was changed via the instructions. Cross-experiment comparisons at first suggested that the size of the diversity effect might depend on causal structure and base rate information. In contrast, Experiment 4, in which the structure and base rate manipulations were replicated within a single experiment, showed a constant size of the diversity effect across all conditions, and, most challenging for causal-reasoning explanations as opposed to simple matching or discounting strategies, the ratings were hardly affected by the manipulations at all. Only in the case of a high base rate did a structure with short causal chains lead to higher ratings than did a structure with long causal chains. This occurred independent of the symptoms' diversity. Thus, a consistent influence of causal structure or base rate information on the size of the diversity effect could not be shown. Computations with causal Bayes nets revealed that despite the complex interplay of parameters, the

remarkably stable human diversity effect corresponds to a stable normative effect across the explored conditions.

The finding that structure and base rate manipulations did not consistently modulate the diversity effect is at variance with reasoning about alternative causation involving causal processes (e.g., causal models). If causal processes are considered, alternative explanations can be imagined more easily for proximal symptoms that can be caused by a single common cause than for diverse symptoms. If symptoms share more intermediate common causes (in the case of longer causal chains), or if the root cause is less probable (in the case of a decreased base rate), alternatives may be imagined more easily, resulting in an increase of the diversity effect. However, such modifications of the diversity effect were not consistently found. Not even expected modifications of the general level of probability ratings were observed.

Nevertheless, the consistent diversity effect could still have resulted from reasoning about alternative causation if such reasoning was independent of the structure and base rate information: Cued by the presented symptoms, retrieving alternative causations from memory would not involve causal processes. There is strong evidence that humans apply this strategy in reasoning with causal conditionals, in which the number of retrieved alternative explanations (counterexamples) can be negatively related to the inferred acceptance of the root cause as the true cause (suppression effect; De Neys et al., 2003). Thus, participants may well have rated the probability of the proposed diagnosis in the present experiments as higher for diverse symptoms because they retrieved fewer or no alternative causes from memory. However, mere memory retrieval of alternative causes does not explain the observed effects of the varied numbers of symptoms, the observed effects of base rate manipulations, and earlier findings with artificial materials.

First, with regard to the varied numbers of symptoms, alternative causations are less probable when sets of three instead of two symptoms point to the root cause. Accordingly, symptom sets that included an unspecific symptom received higher probability ratings, suggesting that unspecific symptoms strengthened the belief in the root cause by decreasing the subjective probability of alternative causations. Moreover, because the chance of alternative causation is especially decreased for proximal sets, the diversity effect for sets of three symptoms should have been smaller than for sets of two. This, however, could only be shown in Experiment 3. Hence, the variation of the number of symptoms provided only marginal evidence for the consideration of alternative causation. Second, if a majority of participants used this strategy, they apparently were not consistently influenced by the base rate manipulation that should have changed the accessibility of alternative causes in memory. Finally, the memory-based strategy does not predict a diversity effect with artificial materials, for which no alternative causes are retrievable in memory. Thus, the diversity effect obtained with artificial materials in a previous experiment (Kim & Keil, 2003) proves that this effect can result not only from memory retrieval.

Alternative ways in which the diversity effect could arise without considering causal processes still cannot be excluded (Kim & Keil, 2003). Two reasoning heuristics that could produce a causal diversity effect in the diagnosis-rating task do not predict a variation by structure or base rate. These heuristics do not require a consideration of the causal processes that lead to the observed symptoms, and do not even involve alternative causation. That is, simply counting the number of supported chains, as well as discounting symptoms that share an intermediate cause as their immediate parent, would not be affected by a change in chain length or base rate changes.

The literature on causal reasoning suggests that in principle, instead of applying the alternative-causation heuristic or even simpler heuristics, reasoners could also have represented alternatives in mental models of the causal structure they had acquired, and might have integrated evidence over these alternatives. For instance, according to one line of research on conditional reasoning, mental models can be tagged with probability information (e.g., Barrouillet, Gauffroy, & Lecas, 2008; Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999). Some research suggests that for causal conditionals, those mental models could be mental causal models (Ali, Chater, & Oaksford, 2011; Fernbach & Erb, 2013). Accordingly, causal reasoning with mental causal models of the acquired causal structure that allows for the incorporation of probabilistic evidence (Krynski & Tenenbaum, 2007; Meder et al., 2014; Waldmann & Hagmayer, 2013) could produce the diversity effect. Thus, the diversity effect remains a suitable focus for studies of causal reasoning with causal models.

Although the scaled response format of probability ratings may have promoted a probabilistic mode of processing in our experiments, as has been shown in conditional reasoning (Markovits, Forgues, & Brunet, 2010), the causal structures and the causal scenario were only qualitatively described to our participants. With quantitatively specified materials, the present task possibly could reveal whether participants use knowledge structures that approximate causal models, or whether they apply cognitive shortcuts that provide approximations of causal judgments (Fernbach & Rehder, 2013). For instance, manipulating the number of alternative explanations within an artificial setting should directly test the influence of cued retrieval of alternative causes on the diversity effect. With a quantitatively specified task and stronger manipulations of base rate and causal structure, the causal-diversity effect might be a useful tool to evaluate general theories of estimating conditional probabilities that can be applied to diagnostic reasoning (Meder et al., 2014), such as power PC theory (Cheng, 1997) or models of causal attribution (Holyoak, Lee, & Lu, 2010).

The diversity effect in diagnostic reasoning proved remarkably stable across multiple conditions.

Principally, reasoning about symptom constellations could be independent from the means of their causation. The present results as the original study cannot exclude simple heuristics that do not consider alternative causations. However, structured causal knowledge was acquired and could have directed reasoning about alternative causation. Because relations between causes and effects can systematically affect the judgments in diagnostic reasoning (Meder, Mayrhofer, & Waldmann, 2009), the diversity effect in diagnostic reasoning remains instrumental for testing the extent to which causal representations contribute to diagnostic judgments.

Author note We thank Ramona Groß and Agnes Scholz, who helped to manage the experiments, and Katharina Behrendt, Theresa Beuster, Susann Geller, Miriam Müller-Bardorff, Maria Stephan, and Barbara Wulfken for their help in conducting pretests and collecting data. This research was supported by German Research Foundation (DFG) Grant Numbers JA 1761/7-1 and KR 1057/17-1.

References

Ali, N., Chater, N., & Oaksford, M. (2011). The mental representation of causal conditional reasoning: Mental models or causal models. *Cognition, 119,* 403–418. doi:10.1016/j.cognition.2011.02.005

Barrouillet, P., Gauffroy, C., & Lecas, J. F. (2008). Mental models and the suppositional account of conditionals. *Psychological Review, 115,* 760–772. doi:10.1037/0033-295X.115.3.760

Byrne, R. M. J. (1989). Suppressing valid inferences with conditionals. *Cognition, 31,* 61–83. doi:10.1016/0010-0277(89)90018-8

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104,* 367–405. doi:10.1037/0033-295X.99.2.365

De Neys, W., Schaeken, W., & D'Ydewalle, G. (2003). Inference suppression and semantic memory retrieval: Every counterexample counts. *Memory & Cognition, 31,* 581–595. doi:10.3758/BF03196099

Fernbach, P. M., & Erb, C. D. (2013). A quantitative causal model theory of conditional reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1327–1343. doi:10.1037/a0031851

Fernbach, P. M., & Rehder, B. (2013). Cognitive shortcuts in causal inference. *Argument & Computation, 4,* 64–88. doi:10.1080/19462166.2012.682655

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 225–244. doi:10.1037/0096-3445.117.3.227

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive Science, 25,* 565–610. doi:10.1207/s15516709cog2504_3

Heit, E., & Hahn, U. (2001). Diversity-based reasoning in children. *Cognitive Psychology, 43,* 243–273. doi:10.1006/cogp.2001.0757

Heit, E., Hahn, U., & Feeney, A. (2005). Defending diversity. In W. Ahn, B. C. Goldstone, A. B. Love, & P. Wolff (Eds.), *Categorization inside and outside of the lab: Festschrift in Honor of Douglas L. Medin* (pp. 87–100). Washington, DC: American Psychological Association.

Holyoak, K. J., Lee, H. S., & Lu, H. (2010). Analogical and category-based inference: A theoretical integration with Bayesian causal models. *Journal of Experimental Psychology: General, 139,* 702–727. doi:10.1037/a0020488

Homa, D., & Vosburgh, R. (1976). Category breadth and the abstraction of prototypical information. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 322–330. doi:10.1037/0278-7393.2.3.322

Jahn, G., & Braatz, J. (2014). Memory indexing of sequential symptom processing in diagnostic reasoning. *Cognitive Psychology, 68,* 59–97. doi:10.1016/j.cogpsych.2013.11.002

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J.-P. (1999). Naïve probability: A mental model theory of extensional reasoning. *Psychological Review, 106,* 62–88. doi:10.1037/0033-295X.106.1.62

Juhos, C., Quelhas, A. C., & Byrne, R. M. (2015). Reasoning about intentions: Counterexamples to reasons for actions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41,* 55–76. doi:10.1037/a0037274

Kim, N. S., & Keil, F. C. (2003). From symptoms to causes: Diversity effects in diagnostic reasoning. *Memory & Cognition, 31,* 155–165. doi:10.3758/BF03196090

Kim, N. S., Yopchick, J. E., & de Kwaadsteniet, L. (2008). Causal diversity effects in information seeking. *Psychonomic Bulletin & Review, 15,* 81–88. doi:10.3758/PBR.15.1.81

Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General, 136,* 430–450. doi:10.1037/0096-3445.136.3.430

Lo, Y., Sides, A., Rozelle, J., & Osherson, D. (2002). Evidential diversity and premise probability in young children's inductive judgment. *Cognitive Science, 16,* 181–206. doi:10.1207/s15516709cog2602_2

López, A. (1995). The diversity principle in the testing of arguments. *Memory & Cognition, 23,* 374–382. doi:10.3758/BF03197238

López, F. J., Cobos, P. L., & Caño, A. (2005). Associative and causal reasoning accounts of causal induction: Symmetries and asymmetries in predictive and diagnostic inferences. *Memory & Cognition, 33,* 1388–1398. doi:10.3758/BF03193371

Markovits, H., Forgues, H. L., & Brunet, M. L. (2010). Conditional reasoning, frequency of counterexamples, and the effect of response modality. *Memory & Cognition, 38,* 485–492. doi:10.3758/MC.38.4.485

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2008). Inferring interventional predictions from observational learning data. *Psychonomic Bulletin & Review, 15,* 75–80. doi:10.3758/PBR.15.1.75

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2009a). The role of learning data in causal reasoning about observations and interventions. *Memory & Cognition, 37,* 249–264. doi:10.3758/MC.37.3.249

Meder, B., & Mayrhofer, R. (2013). Sequential diagnostic reasoning with verbal information. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1014–1019). Austin, TX: Cognitive Science Society.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2009b). A rational model of elemental diagnostic inference. In N. Taatgen, H. van Rijn, L. Schomaker, & J. Nerbonne (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2176–2181). Austin, TX: Cognitive Science Society.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review, 121,* 277–301.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. L. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review, 10,* 517–532. doi:10.3758/BF03196515

Mehlhorn, K., Taatgen, N. A., Lebiere, C., & Krems, J. F. (2011). Memory activation and the availability of explanations in sequential diagnostic reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 1391–1411.

Murphy, K. (2013). Software packages for graphical models. Retrieved from www.cs.ubc.ca/~murphyk/Software/bnsoft.html

Osherson, D., Smith, E. E., Wilkie, O., & López, A. (1990). Category-based induction. *Psychological Review, 97,* 185–200. doi:10.1037/0033-295X.97.2.185

Pearl, J. (2000). *Causality: Models, reasoning and inference* (Vol. 29). Cambridge, MA: MIT Press.

Read, S. J., & Marcus-Newhall, A. R. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology, 65,* 429–447. doi:10.1037/0022-3514.65.3.429

Rebitschek, F. G., Bocklisch, F., Scholz, A., Krems, J. F., & Jahn, G. (2015). Biased processing of ambiguous symptoms favors the initially leading hypothesis in sequential diagnostic reasoning. *Experimental Psychology, 62,* 287–305. doi:10.1027/1618-3169/a000298

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

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

Waldmann, M. R., & Hagmayer, Y. (2013). Causal reasoning. In D. Reisberg (Ed.), *Oxford handbook of cognitive psychology* (pp. 733–752). New York, NY: Oxford University Press. doi:10.1093/oxfordhb/9780195376746.013.0046

World Health Organization. (2013). Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified. In WHO Collaborating Centres for Classification of Diseases, *International statistical classification of diseases and related health problems—Tenth revision* (*ICD-10)*. Geneva, Switzerland: World Health Organization. Retrieved from http://apps.who.int/classifications/icd10/browse/2010/en#/R50-R69