Cognitive Psychology 96 (2017) 54–84

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

Diagnostic causal reasoning with verbal information ☆

Björn Meder [a,*,1], Ralf Mayrhofer [b,1]

[a] Max Planck Institute for Human Development, Center for Adaptive Behavior and Cognition, Lentzeallee 94, 14195 Berlin, Germany
[b] Department of Psychology, University of Göttingen, Gosslerstraße 14, 37075 Göttingen, Germany

ARTICLE INFO

Article history:
Accepted 25 May 2017

Keywords:
Diagnostic reasoning
Causal reasoning
Bayesian models
Order effects
Verbal uncertainty expressions
Verbal reasoning

ABSTRACT

In diagnostic causal reasoning, the goal is to infer the probability of causes from one or multiple observed effects. Typically, studies investigating such tasks provide subjects with precise quantitative information regarding the strength of the relations between causes and effects or sample data from which the relevant quantities can be learned. By contrast, we sought to examine people's inferences when causal information is communicated through qualitative, rather vague verbal expressions (e.g., "*X* occasionally causes *A*"). We conducted three experiments using a sequential diagnostic inference task, where multiple pieces of evidence were obtained one after the other. Quantitative predictions of different probabilistic models were derived using the numerical equivalents of the verbal terms, taken from an unrelated study with different subjects. We present a novel Bayesian model that allows for incorporating the temporal weighting of information in sequential diagnostic reasoning, which can be used to model both primacy and recency effects. On the basis of 19,848 judgments from 292 subjects, we found a remarkably close correspondence between the diagnostic inferences made by subjects who received only verbal information and those of a matched control group to whom information was presented numerically. Whether information was conveyed through verbal terms or numerical estimates, diagnostic judgments closely resembled the posterior probabilities entailed by the causes' prior probabilities and the effects' likelihoods. We observed interindividual differences regarding the temporal weighting of evidence in sequential diagnostic reasoning. Our work provides pathways for investigating judgment and decision making with verbal information within a computational modeling framework.

1. Introduction

The present paper investigates diagnostic causal reasoning with verbal expressions. Natural language contains a plethora of verbal terms for expressing various kinds of uncertainty, such as "frequently," "rarely," "likely," and "probably." In many real-world situations, such linguistic terms are used to communicate probability or frequency information, despite (or because of) the apparent lack of precision. A doctor says that "disease *X* frequently causes symptom *A*," a friend remarks that he "lost weight because of exercising often," and the news states that "car accidents are almost never caused by bad weather

☆ Experiment 1 was presented at the 2013 Annual Meeting of the Cognitive Science Society (Meder & Mayrhofer, 2013); Behavioral data can be downloaded from https://osf.io/gc2nd/.

* Corresponding author.

E-mail addresses: meder@mpib-berlin.mpg.de (B. Meder), rmayrho@uni-goettingen.de (R. Mayrhofer).

[1] Björn Meder and Ralf Mayrhofer contributed equally to this research and share first authorship on this paper.

http://dx.doi.org/10.1016/j.cogpsych.2017.05.002
0010-0285/© 2017 Elsevier Inc. All rights reserved.

alone." Oftentimes, people also need to make inferences and decisions based on such rather vague expressions, for instance, because precise quantitative information is not available or not communicated. One example are law suits, where the prosecution or the defense may present several pieces of evidence in a sequential fashion, such as eyewitness reports or forensic analyses. Each datum may speak for or against the defendant, and the sequential nature of the task requires keeping track of the relative plausibility of the hypotheses under consideration, given the evidence obtained so far. Similarly, a doctor may make diagnostic inferences based on a series of symptoms reported by a patient, such as a headache, dizziness, and vomiting. Knowing that a particular disease frequently causes these symptoms, whereas another disease rarely does, will increase the probability of the former, even though no precise numerical estimates may be available for quantifying the inference.

Although verbal uncertainty expressions are ubiquitous, they do not easily fit with computational models of cognition, which usually require numerical input. Most behavioral studies therefore provide subjects with precise quantitative information, which enables researchers to derive predictions from formal models. For instance, causal reasoning studies typically provide subjects with described numerical information, such as percentages or frequencies (e.g., Hayes, Hawkins, Newell, Pasqualino, & Rehder, 2014; Krynski & Tenenbaum, 2007; Rehder & Burnett, 2005) or sample data (e.g., Mayrhofer & Waldmann, 2015; Meder, Mayrhofer, & Waldmann, 2014; Rottman, 2016; Waldmann & Holyoak, 1992). In contrast, we investigated diagnostic causal inferences from effects to causes based on verbal terms and compared human judgments to those of a matched control group receiving precise numerical information. Our research was motivated by the rich literature on how people understand verbal frequency and probability expressions, and the numerical estimates they assign to different terms (for reviews, see Clark, 1990; Mosteller & Youtz, 1990; Teigen & Brun, 2003; Wallsten & Budescu, 1995). For the present studies, the mapping between words and numbers was provided by a study that elicited numerical estimates for several frequency terms (Bocklisch, Bocklisch, & Krems, 2012). This mapping provided the basis for our comparison of diagnostic reasoning with verbal versus numerical information. We also used the numerical equivalents to derive quantitative predictions from different probabilistic models of diagnostic reasoning.

We investigated three key issues. First, can people make sound diagnostic causal inferences with verbal information? Second, how do they perform relative to a matched control group in which subjects are provided with the corresponding numerical information? Third, what model accounts best for people's judgments in sequential diagnostic reasoning, that is, when inferences are based on multiple, sequentially observed pieces of evidence?

To address these questions, we conducted three experiments in which the subjects' task was to infer the probability of a binary cause (chemical *X* vs. chemical *Y*) from three sequentially observed effects (symptoms such as dizziness or headache). Subjects received either numerical information on the relevant quantities (e.g., the likelihoods of effects; e.g., "chemical *X* causes headache in 66% of cases") or only verbal information (e.g., "chemical *X* frequently causes headache"). Experiments 1 and 2 employed different verbal terms to convey the strength of the cause–effect relations, using a uniform prior over the two causes [i.e., $P(X) = P(Y) = 0.5$]. In Experiment 3, we additionally manipulated the prior probability of the two causes and conveyed base rate information through either verbal terms or numerical information.

We compared subjects' diagnostic judgments to different computational models whose predictions were derived from the numerical equivalents of the verbal expressions used (Bocklisch et al., 2012). The simplest model is based on standard Bayesian inference, which can be used to derive the posterior probabilities of the causes given the evidence available at each time step. This approach, however, is not sensitive to the potential temporal dynamics of belief updating (e.g., primacy or recency effects). We therefore developed a novel Bayesian model that allows for a differential weighting of earlier or more recent evidence. This model also enabled us to investigate interindividual differences regarding the temporal weighting of evidence in sequential diagnostic reasoning, both in the aggregate and on a subgroup level.

1.1. Mapping words to numbers

Several studies have investigated how people understand linguistic expressions of uncertainty, with research dating back to at least the 1940s and 1950s (Cliff, 1959; Lichtenstein & Newman, 1967; Simpson, 1944, 1963; Stone & Johnson, 1959; for reviews see Clark, 1990; Mosteller & Youtz, 1990; Teigen & Brun, 2003; Wallsten & Budescu, 1995). Typically, subjects are presented with different verbal frequency or probability terms (e.g., "frequently," "likely") and are asked to assign a numerical estimate to each expression (e.g., a percentage or frequency estimate). Key issues of interest include within-subject and between-subjects stability and variability in numerical estimates (e.g., Brun & Teigen, 1988; Budescu & Wallsten, 1985; Dhami & Wallsten, 2005; Simpson, 1963), influence of context and considered events (e.g., Harris & Corner, 2011; Wallsten, Fillenbaum, & Cox, 1986; Weber & Hilton, 1990), different elicitation methods (e.g., Hamm, 1991; Wallsten, Budescu, Rapoport, Zwick, & Forsyth, 1986; Wallsten, Budescu, & Zwick, 1993), and alternate ways of formally modeling the representation of verbal terms (e.g., Reagan, Mosteller, & Youtz, 1989; Wallsten, Budescu et al., 1986; Zadeh, 1975).

Although research shows that the perceived meaning can vary depending on context, elicitation method, or as a function of individual differences, the literature also indicates a relatively stable understanding of verbal uncertainty terms. For instance, Simpson (1963) compared the numerical estimates elicited for several frequency terms to an earlier study he conducted in 1944 with a different subject sample and found a remarkably close correspondence: "For only one word, *sometimes*, was the difference greater than five percentage points, and in over one-third of the terms the percentages are identical" (p. 149; his emphasis). Mosteller and Youtz (1990) analyzed 52 verbal expressions examined in 20 different studies and concluded that "the studies give similar, though not identical, results for the same expression when sampling and other sources of variability are considered" (p. 3). These findings suggest that, at least to some extent, verbal uncertainty terms are interpreted in a similar fashion across different people.

This is also important with respect to applied issues, such as the development of an empirically grounded codification for mapping verbal expressions to numerical values or ranges of values (Mosteller & Youtz, 1990). This is critical to avoid a mismatch between the intended and perceived meaning of linguistic uncertainty terms. For instance, Budescu, Broomell, and Por (2009; see also Budescu, Por, Broomell, & Smithson, 2014; Harris, Corner, Xu, & Du, 2013) investigated how laypeople assessed the probability terms used in the 2007 report of the Intergovernmental Panel on Climate Change (IPCC; Intergovernmental Panel on Climate Change, 2007). Although the IPCC report provided an explicit codification for the verbal phrases that were used (e.g., stating that the term "unlikely" is used to describe probabilities in the range 0–33%), Budescu and colleagues found that people's intuitive understanding of the terms often substantially differed from the interpretation guidelines. Berry, Knapp, and Raynor (2002) investigated how people understand verbal frequency terms used in medicine information leaflets (see also Ziegler, Hadlak, Mehlbeer, & Konig, 2013). For instance, according to an E.U. publication on summarizing product characteristics (European Union, 2009, p. 16), the frequency of side effects associated with the verbal term "common" is 1–10%. In stark contrast, when laypeople (students) were asked to assign a numerical estimate to this term (in either a percentage or frequency format), the mean judged likelihood was about 44%. This highlights the importance of considering empirical research regarding people's intuitive understanding of different verbal terms to avoid a potential mismatch between the intended and perceived meaning. Similar issues also arise in other contexts, such as legal reasoning based on forensic evidence (Fenton, Neil, & Lagnado, 2012). For instance, the Association of Forensic Science Providers (2009) and the European Network of Forensic Science Institutes (2015) have proposed guidelines for mapping numerical likelihood ratios into verbal qualitative expressions, but research indicates that people do not necessarily understand and use these verbal terms as intended (Martire, Kemp, Sayle, & Newell, 2014; Thompson & Newman, 2015; see also de Keijser & Elffers, 2012).

1.2. Judgment and decision making with verbal information

Whereas there is a rich literature on how people understand different verbal uncertainty terms, few studies have examined how people reason with verbal information or utilize it in decision making. Key questions are which mode of communication (e.g., verbal probability phrases vs. numerical information) people generally prefer (and under what circumstances), and which mode leads to better information processing. With respect to the preferred mode of communication, Wallsten and Budescu (1995) suggested that people's preference for using verbal terms or numerical information depends on the nature of the subject matter and the perceived uncertainty (the "congruence principle"). If the uncertainty is well defined, such as in lotteries with quantifiable random variation, numerical information is preferred. Conversely, verbal terms are preferred when the uncertainty is high and more imprecise, for instance, in political or economic forecasts, because the vagueness of verbal expressions preserves this imprecision. Consistent with this idea, Olson and Budescu (1997) found a stronger preference for using numerical information when communicating uncertainty about precise events (probability that a fair spinner would land on one of two areas, where the relative size of the areas determined the probability) compared to communicating uncertainty about general-knowledge questions, where the uncertainty was assumed to be less quantifiable. Converging evidence comes from a study by Du, Budescu, Shelly, and Omer (2011) on precise versus imprecise numerical formats, in the context of financial forecasts (expected earnings of companies). They found that range forecasts were perceived as more informative, accurate, and credible than point forecasts, consistent with the idea that the less precise range forecasts are better suited to communicating the associated uncertainty. Participants were also sensitive to the level of precision (width of the interval) relative to the available information, consistent with Wallsten and Budescu's (1995) congruence principle.

Another research question is what communication format leads to better performance. Diverging findings have been obtained, with some studies indicating people perform better with numerical than verbal information, whereas other experiments showed no difference between formats or a superior performance with verbal information. Zimmer (1983) suggested that people's judgments are less biased when using verbal expressions rather than numerical estimates, based on findings from different frequency and probability judgment tasks. Erev and Cohen (1990) found that conveyors of information (e.g., experts assessing different uncertain events in basketball games) preferred using verbal terms over numerical estimates, but that decision makers who had to rate the attractiveness of gambles based on these events preferred to receive numerical estimates. This preference, however, did not carry over to performance (monetary payoffs), with little difference between the two formats. Budescu, Weinberg, and Wallsten (1988) investigated decisions based on numerically or verbally presented probabilities, using a within-subjects design. They first determined individuals' subjective mapping of words to numbers, and subsequently presented subjects with different gambles, which subjects' had to bid for or to rate their attractiveness. The results showed a high internal consistency, but also that performance (in terms of expected payoffs) with verbal presentation was not as good as when information was presented graphically or numerically. Rapoport, Wallsten, Erev, and Cohen (1990) used an urn scenario to compare probabilistic reasoning with verbal and numerical probabilities. Also using a within-subjects design, judgments were either given in the form of verbal probability phrases or in a numerical format. In a separate phase, subjects were asked to estimate numerical values for the provided probability phrases. The results demonstrated a high internal consistency in subjects' judgments, regardless of whether they expressed their opinion through verbal phrases or numerically, although verbal judgments were slightly less accurate than numerical probability judgments.

1.3. Goals and scope

The experiments and models presented here extend existing research in several ways. First, to investigate how people reason with verbal information, we used a diagnostic causal reasoning task. Human diagnostic reasoning has been an important topic in cognitive psychology since the 1970s (Gigerenzer & Hoffrage, 1995; Tversky & Kahneman, 1974; for a review, see Meder & Gigerenzer, 2014), but to the best of our knowledge it has never been investigated how people reason diagnostically with verbal information. This is also true for causal reasoning research in general, which usually provides information numerically or presents subjects with frequency information in the form of individual events. Some studies on human causal reasoning that focus on qualitative patterns of causal inference (e.g., Markov violations) have used qualitative verbal terms to describe the relevant causal relations (Mayrhofer & Waldmann, 2015; Rehder, 2014; Rehder & Waldmann, 2017). However, the particular choice of verbal terms has not been empirically informed, and the focus of these studies was not on comparing reasoning with verbal vs. numerical information.

Second, previous research on judgment and decision making with verbal terms has either used within-subjects designs (e.g., Budescu et al., 1988; Rapoport et al., 1990) to explore the internal consistency when using different modes of presenting knowledge, or focused on potential mismatches between proposed guidelines for translating numerical values to verbal expressions and people's intuitive understanding of those terms (e.g., Berry, Knapp, and Raynor, 2002; Budescu, Broomell, & Por, 2009; Martire et al., 2014). Here, we use a different approach: in our studies, the task was to make inferences based on verbal expressions whose numerical equivalents have been elicited in a different study with different subjects (see below for details). This allows us to test for a shared understanding of verbal terms across subjects, beyond assessing their internal consistency. This is of particular interest given the diverging findings in the literature regarding whether people reason better with numerical or verbal information. Third, we use a sequential inference task, in which people were presented with different pieces of evidence and were asked to make probability judgment after each datum. In contrast to typical diagnostic reasoning tasks, in which people only receive a single piece of evidence and are asked for a single probability judgments, this task enables tracking belief updating given a sequence of observations. The sequential nature of our task is similar to the study of Rapoport et al. (1990), although they used a decontextualized urn scenario in which the likelihood of the data (balls drawn from the urn) was determined by the composition of the urn, which was provided numerically (e.g., "Urn A contains 60 red balls and 40 white balls. Urn B contains 80 red balls and 20 white balls"). In contrast, we provide likelihood information verbally (e.g., "Chemical Zyroxan occasionally causes headache"). In addition, we also manipulate the prior probability of the hypotheses and convey this through verbal terms (Experiment 3), whereas previous research always used a uniform prior (i.e., $P(\text{urn A}) = P(\text{urn B}) = 0.5$). Finally, we present a new Bayesian model of sequential diagnostic reasoning, which can be used to model primacy and recency effects within a probabilistic modeling framework. The model can be used to investigate and quantify the temporal weighting of information, on both the aggregate and individual level, and as a function of task characteristics. This is a key theoretical contribution, because the literature documents a variety of order effects in sequential inferences tasks. Often, such order effects are assessed relative to a normative benchmark, such as Bayes's rule (e.g., Rapoport et al., 1990; Zimmer, 1983), but there are only few attempts to model and quantify such effects (e.g., Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). We conclude by outlining pathways for modeling the vagueness of verbal terms through probability distributions that can be used as input to probabilistic models of cognition, thereby connecting hitherto separate lines of research.

1.4. Diagnostic causal reasoning with verbal information

To investigate how people make probabilistic causal inferences we used a sequential diagnostic reasoning task, in which people were sequentially presented with evidence (symptoms of patients) and had to judge the relative probability of a binary cause event (chemical X vs. Y) after each obtained datum. A key issue for our research is how to derive quantitative predictions from computational models for reasoning tasks with verbal information. Whereas previous research has focused on within-subjects designs, we used the numerical equivalents of frequency expressions from a study by Bocklisch et al. (2012). They asked subjects to provide numerical estimates for 11 (German) verbal expressions (the translation of German to English expressions follows Bocklisch et al.). For instance, given the statement "It is *sometimes* necessary to work at a rapid pace" subjects were asked to provide a numerical estimate that in their opinion best represented the verbal term. Judgments were given in a frequency format (e.g., "in X of 100 work tasks/cases"). (Bocklisch et al. also asked for minimal and maximal values; here we focus on the elicited typical value.) This mapping of words to numbers provided the basis for our empirical studies and for deriving quantitative model predictions. Table 1 shows the numerical values from Bocklisch et al. as well as the corresponding numerical values from the review by Mosteller and Youtz (1990), who analyzed 20 different studies from the literature. The comparison shows that the estimates correspond closely to each other ($r = 0.993$, root-mean-squared error [$RMSE$] = 4.990).

In our experiments, subjects received either verbal or numerical information and were asked to make diagnostic judgments for different symptom combinations. In each trial, three effects were presented sequentially. After each datum, subjects provided a probability estimate regarding the two possible causes. For each time step, we used the numerical estimates associated with the verbal terms to derive posterior probabilities of the causes given the observed effects. The key question was how closely subjects' diagnostic judgments would track the probabilities entailed by the statistical structure of the task.

Table 1
Descriptive statistics for numerical estimates of 11 verbal frequency expressions from the literature.

Frequency expression (original German)	Bocklisch et al. (2012)		Mosteller and Youtz (1990)	
	Mean	*SD*	Mean	Weighted Mean
Never (nie)	1.37	2.23	1	1
Almost never (fast nie)	8.31	5.03	4	3
Infrequent(ly) (selten)	18.52	6.36	17	17
Occasionally (gelegentlich)	28.92	12.23	22	22
Sometimes (manchmal)	33.13	10.96	28	26
In half of the cases (in der Hälfte der Fälle)	50.14	1.21	50	50
Frequent(ly) (häufig)	66.11	15.43	55	61
Often (oft)	69.66	12.91	65	69
Most of the time (meistens)	75.46	9.05	–	–
Almost always (fast immer)	88.11	9.46	91	91
Always (immer)	97.46	6.17	98	99

Note. Words in parentheses denote the corresponding German terms used by Bocklisch et al. (2012) and our studies. Mosteller and Youtz (1990) reported mean values for the expression "as often as not," which we consider equivalent to "in half of the cases;" they did not report values for the expression "most of the time" ("meistens"). The weighted mean denotes the average across studies, weighted by the number of respondents in each study. SD = standard deviation.

Importantly, subjects in the linguistic conditions never received any quantitative, numerical information. In Experiment 1, we used the four verbal expressions "infrequently," "occasionally," "frequently," and "almost always" to convey the strength of the relations between causes (chemical substances) and effects (symptoms).[2] The corresponding numerical mean estimates were 19%, 29%, 66%, and 88%. Diagnostic probability judgments were obtained for six unique symptom sequences, each of which consisted of three sequentially observed symptoms. In Experiment 2, we used the four verbal terms "almost never," "sometimes," "often," and "most of the time"; the corresponding numerical mean estimates were 8%, 33%, 70%, and 75%. Probability judgments for 12 unique symptom sequences were elicited. In Experiment 3 we used six different terms: two for conveying the prior probability (base rate) of the cause events and four for communicating the strength of the cause–effect relations. Twenty-four symptom sequences were used to elicit diagnostic probabilities, combined with two different base rates.

2. Modeling sequential diagnostic causal reasoning

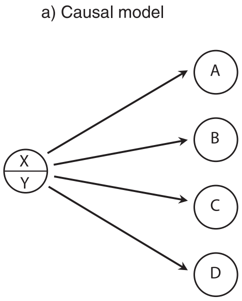
Fig. 1a shows the causal structure underlying the task used in our studies. There are two (mutually exclusive) cause events, *X* and *Y*; each of which can generate effects *A*, *B*, *C*, and *D* probabilistically. The cause variable represented two chemical substances and the effects were different symptoms such as headache and fever. Subjects observed the symptoms sequentially and at each time step the task was to make a diagnostic inference regarding the posterior probabilities of the causes.

In the following, we discuss how sequential diagnostic inferences from effects to causes can be modeled. A common account for modeling belief updating in diagnostic causal reasoning is Bayes's rule (Fernbach, Darlow, & Sloman, 2010; Fernbach, Darlow, & Sloman, 2011; Gigerenzer & Hoffrage, 1995; Hayes et al., 2014; Krynski & Tenenbaum, 2007; Meder & Mayrhofer, 2013; Meder & Mayrhofer, 2017; Meder et al., 2014). However, standard Bayesian inference is not sensitive to the potential temporal dynamics of sequential belief updating, such as primacy and recency effects. There are several studies documenting that diagnostic inferences can be influenced by the order in which information is obtained, such when two pieces of information, *A* and *B*, lead to a different judgment when obtained in the order *AB* versus in the order *BA* (for a review, see Hogarth & Einhorn, 1992). Such effects have been demonstrated in several domains and tasks, including medical diagnosis (Bergus, Chapman, Levy, Ely, & Oppliger, 1998; Chapman, Bergus, & Elstein, 1996), judicial decision making (Kerstholt & Jackson, 1998), and causal reasoning (Rebitschek, Bocklisch, Scholz, Krems, & Jahn, 2015). The order of events does not matter for standard Bayesian inference, but different models can be implemented within a probabilistic modeling framework to account for such effects. We here present a novel Bayesian model of diagnostic reasoning that allows for a differential weighting of earlier or more recent evidence and can be used to model and quantify both primacy and recency effects in sequential diagnostic inference (see Steyvers et al., 2003, for a simpler version that models only recency effects).

2.1. Standard Bayes model

Let $S = \{S_1, \ldots, S_T\}$, where T is the total number of symptoms observed so far, denote a set of symptoms, and let X and Y denote two mutually exclusive causes that can generate S. Since X and Y are mutually exclusive, $P(Y|S) = 1 - P(X|S)$, the posterior probability of cause Y given the symptoms, $P(Y|S)$, can be computed using Bayes's rule:

[2] Because our study was conducted in Germany we used the corresponding German words "selten," "gelegentlich," "häufig," and "fast immer." Note that Bocklisch et al.'s (2012) study was also conducted in Germany with estimates given for the very same (German) terms. We also used the German terms in Experiments 2 and 3.



b) Experiment 1: Relations and presentation formats

Likelihood	Numerical	Verbal
$P(A\|X)$	88%	Almost always
$P(A\|Y)$	19%	Infrequently
$P(B\|X)$	66%	Frequently
$P(B\|Y)$	29%	Occasionally
$P(C\|X)$	29%	Occasionally
$P(C\|Y)$	66%	Frequently
$P(D\|X)$	19%	Infrequently
$P(D\|Y)$	88%	Almost always

c) Experiment 2: Relations and presentation formats

Likelihood	Numerical	Verbal
$P(A\|X)$	75%	Most of the time
$P(A\|Y)$	8%	Almost never
$P(B\|X)$	70%	Often
$P(B\|Y)$	33%	Sometimes
$P(C\|X)$	33%	Sometimes
$P(C\|Y)$	70%	Often
$P(D\|X)$	8%	Almost never
$P(D\|Y)$	75%	Most of the time

Fig. 1. Task design. (a) Causal structure underlying the sequential diagnosis task. (b) Strength of the individual symptom likelihoods in the numerical and verbal formats used in Experiments 1. (c) Strength of the individual symptom likelihoods in the numerical and verbal formats used in Experiment 2. In both experiments subjects were instructed that the two mutually exclusive cause events had equal prior probability, $P(X) = P(Y) = 0.5$.

$$P(Y|S) = \frac{P(S|Y) \cdot P(Y)}{P(S|Y) \cdot P(Y) + P(S|X) \cdot P(X)} \tag{1}$$

where $P(S|Y)$ denotes the likelihood of the symptoms given cause Y, $P(Y)$ is the prior probability of cause Y, and $P(S|X)$ and $P(X)$ denote the corresponding estimates for the alternative cause X. Thus, the posterior probability of cause Y is a function of the likelihood of the set of symptoms S given each of the two causes X and Y and their prior probability.

Fig. 1a provides a graphical illustration of the causal structure underlying the diagnostic reasoning task. The numerical values shown in Fig. 1b are the individual likelihoods with which each of the causes X and Y generated the symptoms in Experiment 1. For instance, X caused symptom A with a likelihood of 0.88, whereas Y caused symptom A with a likelihood of 0.19. Now consider the three-symptom trial sequence *A–C–D*. To infer the probability of the causes given the observed symptoms, Bayes's rule can be applied in a stepwise fashion, with the posterior at time step t serving as the prior for time step $t + 1$. After observing the first symptom A, the posterior probability is $P(Y|A) = 0.18$ (and, therefore, $P(X|A) = 0.82$). This posterior then serves as prior in the next time step when symptom C is observed, yielding $P(Y|A, C) = 0.33$, and so on (cf. Table 2).

Alternatively, the posterior probability can be derived by conditioning on all symptoms at once, using the initial prior over the cause events. In this case, the joint likelihood [e.g., $P(A, D|Y)$] first needs to be inferred from the individual marginal likelihoods $P(A|Y)$ and $P(C|Y)$. Assuming that the symptoms are conditionally independent given the causes (i.e., assuming that the causal Markov condition holds; Pearl, 2000; see also Domingos & Pazzani, 1997; Jarecki, Meder, & Nelson, 2013), the joint likelihood of a set of symptoms is given by multiplying the individual likelihoods separately for each hypothesis (i.e., cause); then Eq. (1) can be applied as usual. Both the sequential updating and the joint updating yield the same posterior probabilities, regardless of the order in which the evidence is obtained. Since all symptoms contribute equally to the posterior in proportion to their likelihoods, we refer to this account as the *standard Bayes model*. Appendix A provides a numerical example for the symptom sequence *A–C–D*.

2.2. Temporal weighting of evidence: "temporal Bayes"

For the standard Bayes model, the temporal order in which evidence is obtained does not matter: The posterior probability is the same regardless of whether beliefs are updated sequentially according to the individual symptoms (with the posterior at time step t serving as prior at time step $t + 1$) or conditional on all symptoms at once.

How can the temporal dynamics of sequential belief updating be modeled? For instance, diagnostic inferences could be influenced by memory limitations, such as the (partial) neglect of earlier obtained evidence in favor of more recent evidence (recency effects). Conversely, diagnostic inferences could be influenced by an overweighting of earlier evidence, relative to

Table 2
Test trials with sequentially presented symptoms in Experiment 1.

Posterior probability	Symptom sequence					
	A–D–C	*D–A–C*	*B–C–A*	*C–B–A*	*A–C–D*	*C–A–D*
$P(Y\|S_1)$	0.18	0.82	0.31	0.69	0.18	0.69
$P(Y\|S_1, S_2)$	0.50	0.50	0.50	0.50	0.33	0.33
$P(Y\|S_1, S_2, S_3)$	0.69	0.69	0.18	0.18	0.69	0.69

Note. Numbers refer to the posterior probability of cause Y given a set of symptoms $S_i \in \{A, B, C, D\}$ according to the standard Bayes model, based on the numerical likelihoods in Fig. 1b and a uniform prior over the cause events, $P(X) = P(Y) = 0.5$.

later obtained evidence (primacy effects). To model the (potential) influence of time in sequential diagnostic reasoning, we applied the log odds form of Bayes's rule to the target inference:

$$\varphi = \log\frac{P(Y|S)}{P(X|S)} = \log\frac{P(Y)}{P(X)} + \log\frac{P(S|Y)}{P(S|X)} \tag{2}$$

where the first summand is the prior odds and the second summand is the likelihood odds. [If the two causes X and Y are equiprobable a priori, $P(X) = P(Y) = 0.5$, the log prior odds can be omitted from Eq. (2)]. Assuming that symptoms are conditionally independent given each cause we can expand the likelihood odds by summing over the sequence of symptoms S_1, . . ., S_T:

$$\varphi = \log\frac{P(Y|S)}{P(X|S)} = \log\frac{P(Y)}{P(X)} + \sum_{t=1}^{T}\log\frac{P(S_t|Y)}{P(S_t|X)} \tag{3}$$

where t is the current symptom and T is the total number of symptoms observed so far. The log posterior odds can then be transformed into a conditional probability by an inverse-logit transformation:

$$P(Y|S) = \frac{1}{1+e^{-\varphi}} \tag{4}$$

Eqs. (3) and (4) are mathematically equivalent to the standard form of Bayes's rule for the posterior probability of cause Y given the set of symptoms S as shown in Eq. (1) (see Appendix A for a numerical example for the sequence A–C–D).

The log-odds form allows us to introduce an exponential weighting parameter δ that controls the weighting of symptoms as a function of the temporal order in which they are observed (cf. Steyvers et al., 2003). The weighting function $w_\delta(t)$ for a given δ and T is given by

$$w_\delta(t) = \begin{cases} e^{\delta(t-T)} & \text{if } \delta > 0 \\ e^0 = 1 & \text{if } \delta = 0 \\ e^{\delta(t-1)} & \text{if } \delta < 0 \end{cases} \tag{5}$$

where t is the current symptom and T is the total number of symptoms observed so far. Thus, $w_\delta(t)$ is a continuous function over δ. Given the set of weights over the symptoms, the log posterior odds are given by

$$\varphi = \log\frac{P(Y)}{P(X)} + \sum_{t=1}^{T} w_\delta(t)\cdot\log\frac{P(S_t|Y)}{P(S_t|X)} \tag{6}$$

Using the inverse-logit transformation (Eq. (4)), the log posterior odds can then be transformed into the target quantity $P(Y|S)$. Depending on the value of the exponential weighting parameter δ, the symptoms are weighted differently when deriving the posterior probability (see Appendix A for a numerical example).

Our *temporal Bayes model* satisfies three constraints. First, it contains the standard Bayes model as a special case when δ = 0; in this case each symptom is weighted equally (i.e., contributes to the sum of log likelihood odds in Eq. (3) according to the symptom's likelihood). Also, for the first piece of evidence (or if there is only a single datum available), our model is equivalent to the standard Bayes model regardless of the value of δ. Second, if δ > 0, more weight is placed on more recently obtained evidence, meaning that earlier evidence has less influence on the posterior probability. In the limit, as $\delta \to \infty$, only the last observed symptom is considered, which corresponds to an inference strategy that is completely ignorant of past information (e.g., a diagnostic reasoner without memory). Third, if δ < 0, more weight is placed on earlier evidence, relative to more recent observations. In the limit, as $\delta \to -\infty$, the log likelihood odds are completely determined by the likelihoods of the symptom that was observed first, ignoring all subsequent evidence.

Fig. 2 illustrates the weighting function $w_\delta(t)$ for different values of $\delta \in \{-10, -1, 0, 1, 10\}$ for three sequentially observed symptoms. The first row shows the symptom weights for the first symptom. In this case, regardless of the value of δ, the weight attached to the log likelihood ratio is 1; therefore the inferred posterior probability is always identical to the predictions of the standard Bayes model. The second row shows the weighting function for two sequentially observed symptoms. If δ = 0, both symptoms are weighted equally, as in standard Bayesian inference. For negative values of δ, the first symptom receives a higher weight than the second symptom. As shown below, this yields a primacy effect as the initially observed evidence has a stronger influence on the posterior probability than the later symptom. The plot where δ = −10 illustrates the extreme case: The weight of the second symptom is close to zero, in which case the posterior probability is almost exclusively determined by the first symptom. Conversely, for positive values of δ the first symptom receives *less* weight than the second observed symptom. For a value of δ = 10, the weight of the initially observed symptom is close to zero, entailing that the second symptom almost exclusively determines the posterior probability. The third row shows the weighting functions for a three-symptom sequence. Again, for δ = 0 all symptoms are equally weighted. For negative values of δ, earlier symptoms receive a higher weight than later symptoms; for positive values of δ, later symptoms receive a higher weight than earlier symptoms. The limiting cases are illustrated by δ = −10 and δ = 10; in these cases, almost all the weight is given to either only the first symptom or only the last symptom, respectively.

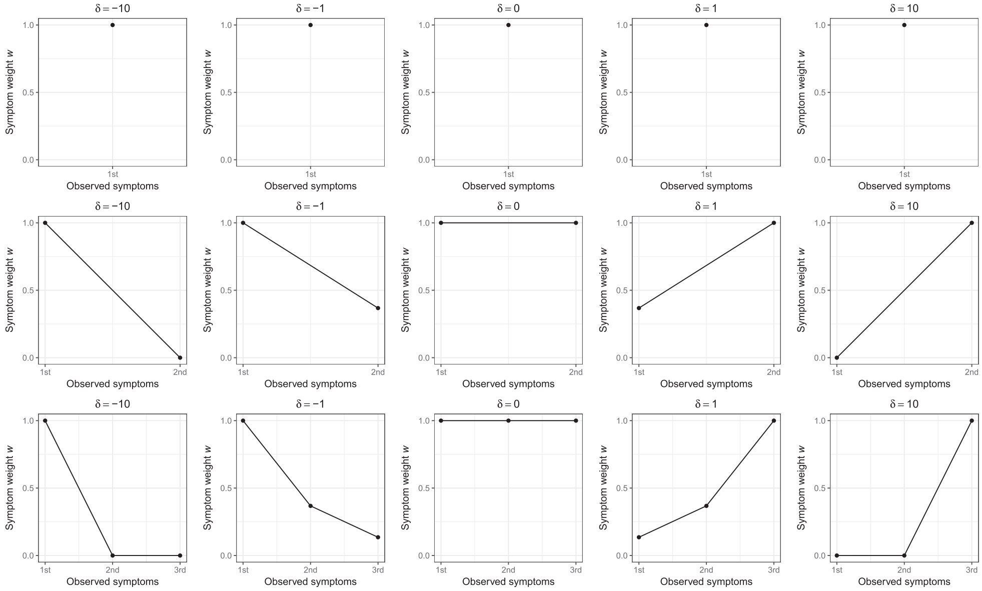


Fig. 2. Weighting function $w_\delta(t)$ of the temporal Bayes model for different values of weighting parameter δ and different numbers of symptoms (one, two, or three). If δ = 0 (middle column), all symptoms are weighted equally. In this case, the predictions of the temporal Bayes model are equivalent to the predictions of the standard Bayes account. If δ < 0, earlier symptoms receive more weight than later symptoms (left two columns, δ = −10 and δ = −1). If δ > 0, later symptoms receive more weight than earlier symptoms (right two columns, δ = 1 and δ = 10).

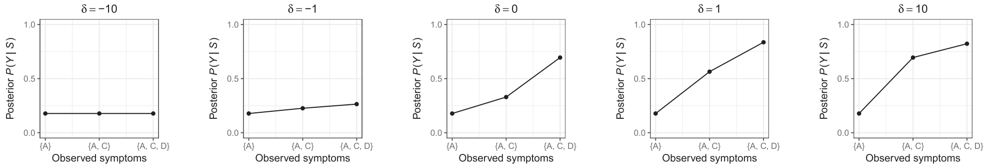


Fig. 3. Influence of the weighting function $w_\delta(t)$ on the inferred posterior probability $P(Y|S)$, for the symptom sequence $S = A–C–D$. If δ = 0, the inferred posterior probabilities correspond to the predictions of the standard Bayes model (middle column). If δ < 0 (two left columns), primacy effects result: Since more weight is placed on earlier observed symptoms, the derived posterior probabilities are more strongly influenced by earlier observed evidence. If δ > 0 (two right columns), recency effects result: Since more weight is placed on later observed symptoms, the derived posterior probabilities are more strongly influenced by more recent evidence.

Fig. 3 illustrates the influence of the weighting function on the target inference $P(Y|S)$ for the symptom sequence $A–C–D$, given a uniform prior over the two causes [$P(X) = P(Y) = 0.5$] and the symptom likelihoods used in Experiment 1 (Fig. 1b). The middle column shows the derived posterior probabilities for δ = 0, which correspond to the predictions of the standard Bayes model. The three posterior probabilities are $P(Y|A) = 0.18$, $P(Y|A, C) = 0.33$, and $P(Y|A, C, D) = 0.69$.

A different set of posterior probabilities results if δ ≠ 0. For negative values of δ, a primacy effect results: If δ = −10, beliefs are updated after the first symptom *A* (going down from the prior probability of 0.5 to 0.18) but remain essentially unaffected by the subsequently observed symptoms *C* and *D*. If δ = −1, symptoms *C* and *D* are not completely ignored, but they have less influence on the posterior probability than the initial symptom *A*. Conversely, for positive values of δ, more recent symptoms have a stronger influence on the target inference than earlier symptoms, modeling a recency effect. For instance, according to the standard Bayes model (δ = 0), the posterior probability after observing all three symptoms is 0.69, whereas it is 0.86 according to the temporal Bayes model with δ = 1 (see Appendix A for a step-by-step derivation). The inference pattern for δ = 10 illustrates the extreme case in which the posterior probability is almost exclusively determined by the currently observed symptom, neglecting all previously acquired evidence.

In summary, our temporal Bayes model can model differential weighting of evidence in sequential diagnostic reasoning. Depending on δ, the weighting function $w_\delta(t)$ places more weight on earlier symptoms, relative to subsequently observed symptoms, or more weight on later symptoms, relative to earlier observed symptoms. This differentially influences the inferred diagnostic probability, eventually resulting in primacy or recency effects.

3. Experiment 1

The main goal of Experiment 1 was to investigate diagnostic causal reasoning with verbal information, relative to a condition in which information was presented numerically and with respect to the standard and temporal Bayes models. In addition, we explored whether and how the way the symptoms are presented when making diagnostic inferences influences the temporal weighting of evidence.

Our main manipulation was the way in which subjects were informed about the strength of the relations between causes and effects. In the verbal condition the strength of the individual causal relations was conveyed through four verbal terms (“infrequently”, “occasionally”, “frequently”, “almost always”). In the numerical condition, causal strengths were presented in a percentage format. The two formats were matched using the estimates from Bocklisch et al. (2012). For instance, in the verbal condition subjects learned that “*X* almost always causes *A*”, whereas in the numerical condition subjects learned that “*X* causes *A* in 88% of cases” (Fig. 1b).

With the second manipulation, we aimed to investigate possible influences of temporal weighting on diagnostic judgments. In the all-symptoms condition, the full set of symptoms reported so far was visible on the screen when subjects made a diagnosis. By contrast, in the single-symptom condition, only the current symptom was visible on the computer screen when subjects made a diagnostic judgment (Fig. 4). Thus, in the all-symptoms condition subjects saw the sequence $\{S_1\}$, $\{S_1, S_2\}$, $\{S_1, S_2, S_3\}$, whereas in the single-symptom condition subjects were presented with the sequence $\{S_1\}$, $\{S_2\}$, $\{S_3\}$. The rationale behind this manipulation was that there might be a tendency to overweight the currently presented symptom (i.e., stronger recency effects compared to the all-symptoms condition).

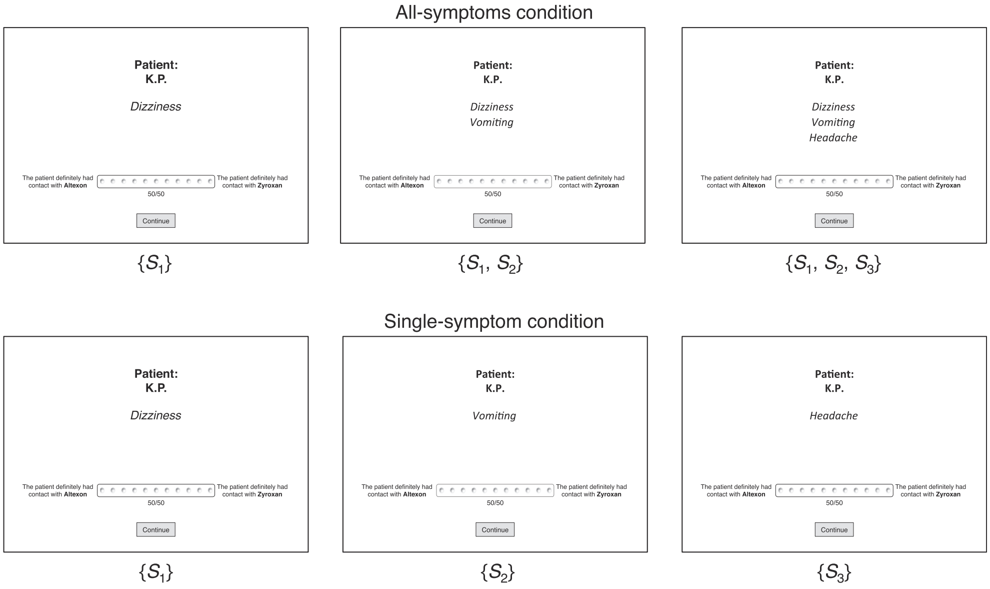
3.1. Method

3.1.1. Subjects and design

One hundred fifty-six students (103 women, $M_{\text{age}} = 23.4$ years; one subject did not provide information on gender and age) from the University of Göttingen, Germany, participated in this experiment as part of a series of various unrelated computer-based experiments. Subjects either received course credit or were paid €8/h; they were randomly assigned to one of the 2 (numerical vs. verbal) × 2 (single vs. all symptoms) = 4 between-subjects conditions. This and the two subsequent experiments were conducted on a computer. Mean completion time in Experiment 1 was 21 min.

3.1.2. Materials and procedure

We used a hypothetical medical diagnosis scenario in which the subjects’ task was to find out which of two fictitious chemical substances was the cause of certain symptoms in patients. We asked subjects to take the role of a doctor being responsible for the workers at two chemical plants. At one plant workers could come in contact with the substance “Altexon”; at the other they could have contact with “Zyroxan” (causes *X* and *Y*; Fig. 1a). Each of these substances causes



Patient:
K.P.
Dizziness
The patient definitely had contact with **Altexon** ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ The patient definitely had contact with **Zyroxan**
50/50
Continue

Patient:
K.P.
Dizziness
Vomiting
The patient definitely had contact with **Altexon** ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ The patient definitely had contact with **Zyroxan**
50/50
Continue

Patient:
K.P.
Dizziness
Vomiting
Headache
The patient definitely had contact with **Altexon** ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ The patient definitely had contact with **Zyroxan**
50/50
Continue

Patient:
K.P.
Dizziness
The patient definitely had contact with **Altexon** ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ The patient definitely had contact with **Zyroxan**
50/50
Continue

Patient:
K.P.
Vomiting
The patient definitely had contact with **Altexon** ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ The patient definitely had contact with **Zyroxan**
50/50
Continue

Patient:
K.P.
Headache
The patient definitely had contact with **Altexon** ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ The patient definitely had contact with **Zyroxan**
50/50
Continue

Fig. 4. Symptom presentation conditions in Experiment 1. In the all-symptoms condition, all evidence obtained so far is present on the screen (symptoms $S = \{S_1, \ldots, S_T\}$). In the single-symptom condition, only the current evidence is presented on the screen (symptom S_t). Experiments 2 and 3 used only the all-symptoms condition.

four symptoms: dizziness, fever, headache, and vomiting. The assignment of symptom labels to effects (*A*, *B*, *C*, *D*; Fig. 1a) was randomized across subjects.

Subjects were informed that their task would be to diagnose a group of workers who had had contact with one of the two substances. The instructions explicitly stated that accidents were equally likely to happen in each of the two plants (i.e., the base rate of each cause was 50%), and that all symptoms were caused by either Altexon or Zyroxan, because the workers came from different plants and immediately went to see the doctor after having come in contact with either substance (i.e., the hypotheses are mutually exclusive and exhaustive). Subjects were also told that the patients would report their symptoms sequentially. To proceed, subjects had to complete an instruction test with four multiple-choice questions. Questions referred to the possible symptoms caused by the chemicals, the names of the chemicals, the (uniform) base rate of the causes, and that the causes were mutually exclusive. All questions had to be answered correctly, otherwise subjects had to reread the instructions and take the test again.

The main part of the experiment consisted of two phases: a learning phase, in which subjects learned the strengths of the individual causal relations, and a diagnostic reasoning phase, in which subjects were sequentially presented with symptoms of different patients and had to make a diagnostic judgment after each symptom. Fig. 1b illustrates the strengths of the causal relations (likelihoods) between substances and symptoms according to the two presentation formats. In the learning phase, the subjects' task was to learn the strength of the individual relations in a trial-by-trial fashion. On each trial, subjects were shown a substance along with a symptom and had to estimate how frequently the substance causes the symptom. In the verbal condition, possible answers were "infrequently," "occasionally," "frequently," and "almost always." In the numerical condition, the corresponding answers were 19%, 29%, 66%, and 88%. After giving a response, subjects received feedback regarding the actual relation. The eight relations were presented blockwise, with the order randomized within each block. To proceed to the diagnostic reasoning phase, subjects needed to answer two consecutive blocks correctly (or pass 12 blocks in total).

In the diagnostic reasoning phase, the subjects' task was to make diagnostic judgments for different sequences of symptoms, with each symptom sequence referring to a different patient who had come in contact with either *X* or *Y*. Each test trial consisted of three sequentially presented symptoms (e.g., *A*–*D*–*C*), with a diagnostic judgment requested after each symptom. In the all-symptoms condition, all symptoms reported so far were present on the screen. In the single-symptom condition, only the current symptom was displayed (Fig. 4). All judgments were given on an 11-point scale from 0 (*The patient definitely had contact with Altexon*) to 100 (*The patient definitely had contact with Zyroxan*) in steps of 10; that is, subjects gave an estimate of $P(Y|S)$. The midpoint of the scale was labeled 50/50. In all experiments, at the beginning of each diagnostic trial, the scale was reset to the midpoint of the scale. Within each trial (symptom sequence), the judgment made on the previous step remained visible.

Table 2 shows the six symptom sequences together with the posterior probabilities derived using the likelihoods shown in Fig. 1b, assuming $P(X) = P(Y) = 0.5$. Additionally, we presented the six symptom sequences that entailed identical posterior probabilities for *X* [e.g., $P(Y|A, D, C) = P(X|D, A, B) = 1 - P(X|A, D, C)$] such that diagnoses were counterbalanced.[3] Thus, each subject saw 12 sequences in total; the corresponding pairs were later recoded and averaged within subjects.

All test trials were administered in random order. After the diagnostic reasoning phase, we tested subjects again with respect to the strength of the individual chemical–symptom relations (as learned in the learning phase) by presenting an additional block of the learning phase (without feedback). This manipulation check served as learning criterion to test whether subjects still remembered the relations between causes (chemicals) and effects (symptoms).

3.2. Results and discussion

3.2.1. Learning criterion

At the end of the experiment, we tested subjects on the eight substance–symptom relations presented in the learning phase. Because the strength of the individual relations was the basis for the diagnostic judgments, we excluded all subjects who could not reproduce at least seven of the eight relations correctly. Accordingly, 28.2% of the subjects were excluded from the analyses, yielding between 27 and 30 valid subjects per condition (total $N = 112$).

3.2.2. Overall analyses

In total, we obtained 112 (subjects) × 12 (symptom sequences) × 3 (sequential judgments) = 4032 diagnostic judgments. For the analyses, we first recoded and aggregated corresponding pairs of trials (see above) within subjects and then averaged across subjects, yielding 6 (symptom sequences) × 3 (judgments) = 18 mean diagnostic judgments per condition (verbal vs. numerical presentation/single symptom vs. all symptom).

Fig. 5 shows subjects' mean diagnostic judgments for the different symptom sequences along with the posterior probabilities derived from the standard Bayes model (note that the standard Bayes model has no free parameters, so no fitting is involved here). A first inspection of the data indicates that subjects' judgments were remarkably consistent, with estimates being close to the derived posteriors regardless of whether information was provided in a verbal or numerical format. The

[3] For instance, with respect to cause *Y*, the symptom sequence *A*–*D*–*C* entails the following posterior probabilities according to the standard Bayes model: $P(Y|A) = 0.18$, $P(Y|A, D) = 0.5$, and $P(Y|A, D, C) = 0.69$. Conversely, with respect to cause *X*, the symptom sequence *D*–*A*–*B* entails the same posterior probabilities: $P(X|D) = 0.18$, $P(X|D, A) = 0.5$, and $P(X|D, A, B) = 0.69$. We therefore aggregated the two corresponding sets of judgments within subjects.

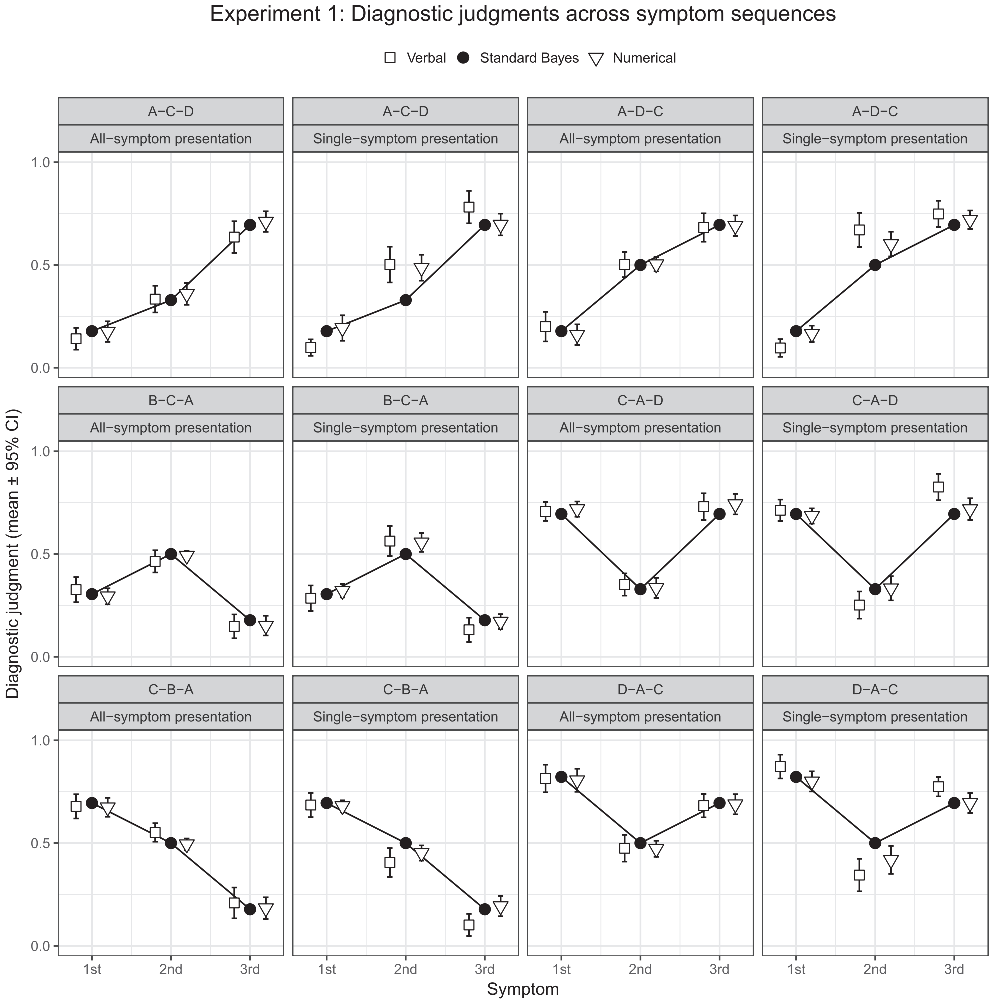


Fig. 5. Mean diagnostic judgments (and 95% confidence intervals) in the verbal and numerical information format conditions in Experiment 1, and predictions of the standard Bayes model. Each plot corresponds to a trial consisting of three sequentially observed symptoms. Judgments shown separately by testing procedure. *A*, *B*, *C*, and *D* are symptoms. All-symptom presentation = all symptoms observed so far presented on screen; single-symptom = only current symptom presented on screen (see Fig. 4).

plots also reveal some influence of the way symptoms were presented during the diagnostic reasoning phase. For instance, for the symptom sequence *A*–*C*–*D*, subjects tended to overestimate the posterior probability of cause *Y* when only the current symptom was displayed (single-symptom condition) compared to in the all-symptom condition, indicating an overweighting of the current symptom (see below for detailed analyses).

Fig. 6a (left) shows the correspondence of the diagnostic judgments in the verbal and numerical condition when aggregating across trials and symptom presentation conditions. Across all obtained diagnostic judgments, the correlation between the two conditions was $r = 0.983$, $RMSE = 0.050$. In addition, judgments closely correspond to the posterior probabilities derived from the standard Bayes model in the numerical ($r = 0.983$, $RMSE = 0.040$) as well as in the verbal condition ($r = 0.962$, $RMSE = 0.070$; see Fig. 6a, middle and right columns).

Taken together, the results demonstrate that subjects were capable of making pretty accurate inferences (with respect to the probabilities derived from the standard Bayes model) when reasoning with verbal information. This performance is particularly remarkable because the numerical equivalents of the verbal terms were taken from another sample of subjects in a different study (Bocklisch et al., 2012). Also, regardless of whether causal strengths were conveyed through numbers or words, subjects' diagnostic judgments quite accurately tracked the posterior probabilities derived from the standard Bayes model across the different symptom sequences.

3.2.3. Model comparison

To evaluate subjects' overall accuracy, we computed the correlation and RMSE between the empirical judgments and the posterior probabilities derived from the standard Bayes model, separately for each of the four conditions. The correlation

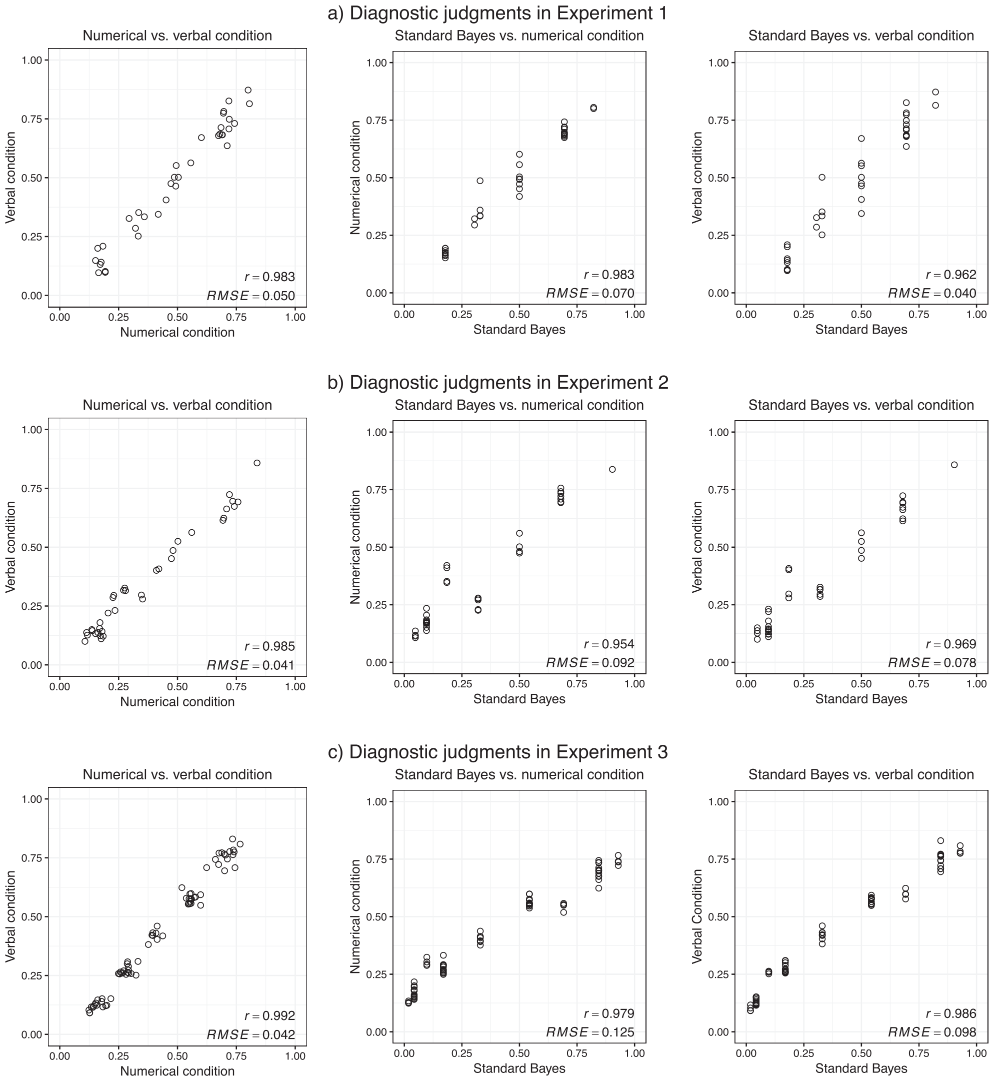


Fig. 6. Scatter plots of mean diagnostic judgments in Experiments 1–3, aggregated over trials and symptom presentation condition. The left column shows the diagnostic judgments in the verbal condition plotted against judgments in the numerical condition. The middle column shows the diagnostic judgments in the numerical condition plotted against the predictions of the standard Bayes model. The right column shows the diagnostic judgments in the verbal condition plotted against the predictions of the standard Bayes model. *r* = Pearson correlation, *RMSE* = Root-mean-squared error.

indicates how well the model predictions can account for the overall trends; the RMSE indicates how well the model accounts for the absolute size of the judgments. To address if symptoms are weighted differently in sequential reasoning, we fitted the weighting parameter δ of our temporal Bayes model to the data (separately for each condition, using the mean-squared error, MSE, as fitting criterion). For this purpose, we used a grid search over a plausible set of values for δ between -10 and +10, in steps of 0.01. The relative size of the decay parameter δ in the single-symptom vs. all-symptoms condition gives an idea of whether the testing procedure influences people's judgments, in particular whether there is a tendency to underweight previous evidence when only the current symptom is presented (i.e., recency effects). This should result in higher values of the weighting parameter δ for the single-symptom condition relative to the all-symptoms condition. (Remember that if $\delta = 0$ each symptom is weighted equally. If $\delta > 0$, more weight is placed on more recent evidence. If $\delta < 0$, more weight is placed on earlier evidence.)

Table 3 shows the fits of the standard Bayes and the temporal Bayes model. Overall, both the (high) correlations and the (low) RMSEs indicate that the models' predictions fit well with subjects' judgments. In the all-symptoms conditions, the fit of the standard Bayes model was almost perfect ($r = 0.991$ and $r = 0.996$, respectively), mirroring that for 35 of 36 (6 trials × 3 symptoms × 2 formats [verbal vs. numerical]) data points, the model's predictions fell inside the 95% confidence interval (CI).

Table 3
Comparison of the standard Bayes and temporal Bayes models in Experiments 1–3.

Format	Symptom presentation	Standard Bayes				Temporal Bayes				
		r	*RMSE*	AIC	BIC	*r*	*RMSE*	δ	AIC	BIC
Experiment 1 (*N* = 112)										
Verbal	All	0.991	0.029	−127.87	−127.87	0.992	0.026	−0.03	−128.73	−127.84
	Single	0.952	0.094	−85.04	−85.04	0.982	0.061	0.58	−98.88	−97.99
Numerical	All	0.996	0.019	−141.96	−141.96	0.997	0.019	0.03	−141.63	−140.74
	Single	0.971	0.053	−105.86	−105.86	0.988	0.036	0.23	−118.14	−117.24
Experiment 2 (*N* = 61)										
Verbal	All	0.969	0.078	−183.49	−183.49	0.980	0.067	−0.13	−192.89	−191.31
Numerical	All	0.954	0.092	−171.51	−171.51	0.960	0.088	−0.09	−173.16	−171.58
Experiment 3 (*N* = 119)										
Verbal	All	0.986	0.098	−334.80	−334.80	0.985	0.095	−0.07	−336.57	−334.29
Numerical	All	0.979	0.125	−299.21	−299.21	0.975	0.120	−0.12	−302.73	−300.46

Note. δ denotes the weighting parameter of the temporal Bayes model, fitted using the mean-squared error as criterion. If $\delta = 0$, the temporal Bayes model correspond to the standard Bayes model, with each symptom being weighted equally. If $\delta > 0$, more weight is placed on more recent evidence; if $\delta < 0$, more weight is placed on earlier evidence. *r* = Pearson correlation, RMSE = root-mean-squared error, AIC = Akaike information criterion, BIC = Bayesian information criterion.

Inspection of the fitted δ parameter of the temporal Bayes model indicates some neglect of previous evidence in both single-symptom conditions, in which only the current symptom was displayed on the screen when subjects made a diagnostic judgment (Fig. 2). Whereas δ was close to zero in the two all-symptoms conditions (−0.03 and 0.03, respectively), diagnostic judgments in the single-symptom conditions were better accounted for by a positive value of the weighting parameter (0.58 and 0.23, respectively). Since a positive δ entails a stronger weight placed on later evidence relative to earlier evidence, this suggests a small recency effect when only the current symptom was presented in the diagnostic reasoning phase (Table 3). Accordingly, for the single-symptom conditions, the temporal Bayes model achieved a higher fit than the standard Bayes model, in terms of both the correlation and the RMSE, whereas both criteria remained essentially unchanged in the all-symptoms conditions. This result indicates that subjects were more likely to neglect previous evidence when it had to be recalled from memory.

We also computed the Akaike information criterion (AIC; Akaike, 1973) and Bayesian information criterion (BIC; Schwarz, 1978) for the standard and temporal Bayes models, separately for each of the four conditions (Table 3).[4] These model-selection criteria take into account that the temporal Bayes model has a free parameter and penalize it accordingly (for details see Burnham & Anderson, 1998; Raftery, 1995; Wagenmakers & Farrell, 2004). Consistent with the above results and the fact that the δ parameter of the temporal Bayes model is close to zero (which corresponds to the standard Bayes model), in both all-symptoms presentation conditions the two models had similar AIC and BIC values. By contrast, in both single-symptom presentation conditions the temporal Bayes model yielded lower AIC and BIC values, indicating that this model provides a better account of the data than the standard Bayes model.

3.2.4. Model-based clustering

Temporal weighting of cumulative evidence might not be due to the task characteristics or the reasoning context alone but might also result from interindividual differences. We therefore explored if it is possible to identify homogeneous subgroups of subjects who differed with respect to their temporal weighting of symptoms (i.e., who differed in the δ parameter). This is also important to rule out possible aggregation artifacts, for instance that the δ values close to zero in the all-symptoms conditions result from aggregating over a bimodal distribution consisting of subjects who either underweight or overweight current evidence.

To identify such clusters, we adapted the model-based clustering technique introduced by Steyvers et al. (2003), which was inspired by *K*-means clustering. The clustering problem requires solving two problems simultaneously: first, assigning subjects to clusters such that clusters are homogeneous with respect to the model predictions, and second, estimating the best fitting δ parameter for each cluster. We defined three qualitatively different clusters a priori: a “primacy cluster” with a to-be-fitted $\delta < 0$, a “standard Bayes cluster” with a fixed $\delta = 0$, and a “recency cluster” with a to-be-fitted $\delta > 0$.

More specifically, the clustering algorithm proceeds as follows: We initialized the model with three clusters defined by $\delta = -0.01$, $\delta = 0$, and $\delta = 0.01$. These values were conservatively chosen so that the initial difference between predictions of the standard Bayes model and the two temporal Bayes versions were small. Predictions for the diagnostic probabilities were generated separately for each cluster and each individual subject was assigned to the cluster with the best corresponding

[4] The AIC was computed according to $n \log(MSE) + 2k$ (Burnham & Anderson, 1998), where n is the number of data points (mean diagnostic judgments), MSE is the mean-squared error (deviations of model predictions from mean judgments), log denotes the natural logarithm, and k is the number of parameters ($k = 0$ for the standard Bayes model and $k = 1$ for the temporal Bayes model). The BIC was computed according to $n \log(MSE) + \log(n)k$. AIC and BIC give identical values for the standard Bayes model, since it has no free parameters.

Table 4
Results of the model-based clustering in Experiments 1–3.

Format	Symptom presentation	Primacy cluster ($\delta < 0$)				Standard Bayes cluster ($\delta = 0$)				Recency cluster ($\delta > 0$)			
		n (%)	δ	*r*	*RMSE*	*n*	δ	*r*	*RMSE*	*n*	δ	*r*	*RMSE*
Experiment 1 (*N* = 112)													
Verbal	All	11 (39%)	– 0.25	0.980	0.042	14 (50%)	0.00	0.989	0.038	3 (11%)	2.50	0.966	0.084
	Single	4 (15%)	– 0.26	0.975	0.052	12 (44%)	0.00	0.984	0.050	11 (41%)	10.00	0.994	0.096
Numerical	All	9 (33%)	– 0.17	0.982	0.053	15 (56%)	0.00	0.994	0.048	3 (11%)	0.79	0.979	0.064
	Single	6 (20%)	– 0.18	0.942	0.078	15 (50%)	0.00	0.996	0.024	9 (30%)	2.13	0.993	0.041
Experiment 2 (*N* = 61)													
Verbal	All	20 (63%)	– 0.21	0.974	0.070	9 (28%)	0.00	0.969	0.070	3 (9%)	0.21	0.895	0.128
Numerical	All	13 (45%)	– 0.20	0.954	0.086	12 (41%)	0.00	0.942	0.098	4 (14%)	0.17	0.931	0.126
Experiment 3 (*N* = 119)													
Verbal	All	23 (38%)	– 0.21	0.975	0.107	23 (38%)	0.00	0.988	0.083	14 (23%)	0.27	0.965	0.122
Numerical	All	27 (46%)	– 0.29	0.974	0.129	28 (47%)	0.00	0.976	0.112	4 (7%)	2.55	0.953	0.109

Note. δ denotes the weighting parameter of the temporal Bayes model, fitted using the mean-squared error as criterion (except for the simple Bayes cluster with $\delta = 0$ being fixed). If $\delta = 0$, the predictions of the temporal Bayes model corresponds to the predictions of the standard Bayes model, with each symptom being weighted equally. If $\delta > 0$, more weight is placed on more recent evidence, and if $\delta \to \infty$, the posterior probability depends only on the current symptom. Conversely, if $\delta < 0$, more weight is placed on earlier evidence, and if $\delta \to -\infty$, the posterior probability depends on only the first symptom observed. *r* = Pearson correlation, RMSE = Root-mean-squared error.

predictions, such that the MSE between the individual judgments and the model predictions was minimized. The algorithm then proceeds as follows: (a) Given the current assignments of subjects to clusters, find the δ parameter for each cluster that minimizes the MSE of the model predictions with respect to the average response profile (i.e., mean diagnostic judgments) of the subjects within each cluster (except for the simple Bayes cluster with $\delta = 0$ being fixed). For this purpose, we used a grid search over δ from -10 to +10, in steps of 0.01. (b) Given the model predictions for the different clusters, reassign subjects to a cluster such that the MSE between the individual response profile and the model prediction is minimized. Iterate through Steps a and b until no subject changes cluster anymore.

We applied this procedure separately to each of the four between-subjects conditions; the results are shown in Table 4. Consistent with the overall high fit of the standard Bayes account (Fig. 6a), in each condition most subjects were assigned to the standard Bayes cluster (i.e., $\delta = 0$). In the single-symptom conditions, however, relatively more subjects were best described by a positive δ (i.e., recency) than in the all-symptoms conditions (35% vs. 11% of subjects, aggregated over presentation format). Also note that for the recency cluster the best fitting δ in the single-symptom conditions were much higher than in the all-symptoms condition (10 vs. 2.5 in the verbal conditions, and 2.13 vs. 0.79 in the numerical condition). Essentially, this means that the overweighting of more recent evidence was more prevalent and more pronounced in the single-symptom conditions; that is, more subjects were assigned to the recency cluster and they were better characterized by a higher δ value. Finally, some subjects showed primacy effects in the sense that they were best described by a negative value of the δ parameter of the temporal Bayes model. Note, however, that the resulting δ values for the primacy cluster are close to zero in all four conditions, indicating only weak primacy effects.

The clustering results clarify the findings obtained by the overall data analysis. First, in all four conditions the standard Bayes cluster was the biggest one. Second, in the two single-symptom conditions more subjects were assigned to the recency cluster than in the corresponding all-symptoms conditions, and the best fitting δ value was much higher. This also explains why the temporal Bayes model achieved a higher fit than the standard Bayes model in these conditions, whereas in the all-symptoms conditions the difference was marginal. Third, a notable proportion of subjects were assigned to the primacy cluster, but the fact that the δ values are relatively close to zero indicates that this effect was less pronounced than the recency effect. Theoretically this makes sense, as presenting only the current symptom when making a diagnostic judgment required subjects to recall previous evidence from memory, thereby potentially influencing diagnostic judgments.

4. Experiment 2

The goal of Experiment 2 was to test the robustness of the findings of Experiment 1. We used different verbal expressions, which were also taken from the study by Bocklisch et al. (2012). Since the main goal was to contrast diagnostic inferences based on verbal versus numerical information, we employed only the all-symptoms presentation condition. We used the all-symptoms condition because the findings of Experiment 1 show that the order effects were less pronounced in this condition. This way we were able to investigate the correspondence of reasoning with verbal versus numerical information without introducing additional factors, such as memory effects or differential weighting of evidence due to the way symptoms were presented in the single-symptom condition of Experiment 1.

Table 5
Test trials with sequentially presented symptoms in Experiment 2.

Posterior probability	Symptom sequence											
	A–D–C	*D–A–C*	*B–C–A*	*C–B–A*	*A–C–D*	*C–A–D*	*A–B–C*	*A–B–D*	*B–A–C*	*B–A–D*	*A–C–B*	*C–A–B*
$P(Y\|S_1)$	0.10	0.90	0.32	0.68	0.10	0.68	0.10	0.10	0.32	0.32	0.10	0.68
$P(Y\|S_1, S_2)$	0.50	0.50	0.50	0.50	0.18	0.18	0.05	0.05	0.05	0.05	0.18	0.18
$P(Y\|S_1, S_2, S_3)$	0.68	0.68	0.10	0.10	0.68	0.68	0.10	0.32	0.10	0.32	0.10	0.10

Note. Numbers refer to the posterior probability of cause *Y* given a set of symptoms $S_i \in \{A, B, C, D\}$ according to the standard Bayes model, based on the likelihoods in Fig. 1c and equal priors for the cause events, $P(X) = P(Y) = 0.5$.

4.1. Method

4.1.1. Subjects and design

Eighty-four students (59 women, 25 men, $M_{\text{age}} = 24.1$ years) from the University of Göttingen participated. They received course credit or were paid €8/h. Subjects were randomly assigned to either the verbal or the numerical format condition. Mean completion time was 26 min.

4.1.2. Materials and procedure

We used the same materials and procedure as in Experiment 1. The key difference was that we used four different verbal expressions and numerical equivalents, which are shown in Fig. 1c. As in the first study, subjects first had to learn the strength of the relations between causes (chemicals) and effects (symptoms), and subsequently had to make a series of sequential diagnostic judgments. To test for a wider range of probabilistic inferences, subjects were presented with all possible 4! = 24 three-symptom sequences (in randomized order). As before, equivalent pairs that entailed the same set of corresponding posterior probabilities [e.g., $P(Y|A, D, C) = 1 - P(Y|D,A, B)$] were recoded and aggregated within each subject. Table 5 shows the 12 unique symptom sequences and the posterior probabilities derived from the standard Bayes model, based on the likelihoods shown in Fig. 1c and $P(Y) = P(X) = 0.5$.

4.2. Results

4.2.1. Learning criterion

As in Experiment 1, after the diagnostic inference phase, subjects were tested for whether they still remembered the symptom likelihoods learned at the beginning of the experiment. We excluded all subjects who could not correctly reproduce the strength of at least seven of the eight causal relations after the diagnostic reasoning phase. Twenty-three of the 84 subjects (27.4%) failed to meet the criterion and were excluded, leaving 61 valid subjects ($n = 32$ in the verbal condition, $n = 29$ in the numerical condition).

4.2.2. Overall analyses

In total, we obtained 61 (subjects) × 24 (symptom sequences) × 3 (sequential judgments) = 4392 diagnostic judgments. For the analyses, we first aggregated corresponding pairs of trials within each subject and then averaged across subjects; yielding 12 (symptom sequences) × 3 (sequential judgments) = 36 mean diagnostic judgments in each condition.

Fig. 7 shows subjects' diagnostic judgments in the numerical and verbal conditions, separately for each symptom sequence. As in Experiment 1, judgments in both conditions were very similar to each other and corresponded closely to the predictions of the standard Bayes model. Notably, even in the cases where subjects' judgments deviated from the predictions of the standard Bayes model (e.g., after the second symptom in the sequences *C–A–B* and *C–A–D*), judgments in both conditions deviated to the same extent.

Fig. 6b illustrates the high correlation of subjects' mean diagnostic judgments in the verbal vs. numerical condition across all trials. As in Experiment 1, we observed a very close correspondence between diagnostic judgments in the numerical and verbal conditions ($r = 0.985$, $RMSE = 0.041$; Fig. 6b, left column). Also, subjects' diagnostic judgments closely resembled the posterior probabilities derived from the standard Bayes model in the numerical condition ($r = 0.954$, $RMSE = 0.092$) as well as in the verbal condition ($r = 0.969$, $RMSE = 0.078$; see Fig. 6b, middle and right column).

Taken together, the results of Experiment 2 corroborate the findings of Experiment 1, showing that subjects were able to make systematic and coherent quantitative judgments even if they never obtained any numerical information on the causal relations. Most importantly, judgments in the numerical and verbal conditions were almost indistinguishable.

4.2.3. Model comparison

As in Experiment 1, we used the correlation and RMSE between judgments and the posterior probabilities to evaluate which model best accounts for the data. Table 3 shows the results for the standard Bayes model and the temporal Bayes model, with the δ parameter of the latter being fitted to the mean diagnostic judgments (using the MSE as criterion). The results show that both models account very well for subjects' inferences, yielding a high correlation and low RMSE. The size

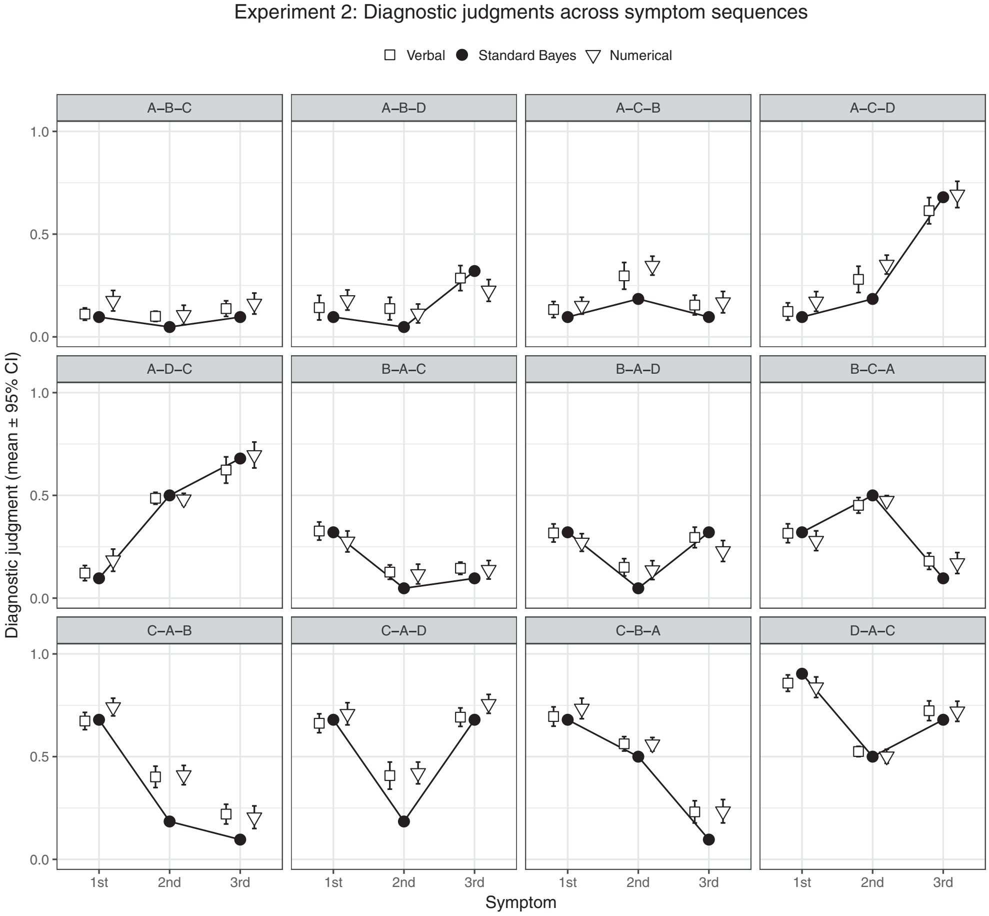


Fig. 7. Mean diagnostic judgments (and 95% confidence intervals) in the verbal and numerical information format conditions in Experiment 2, and predictions of the standard Bayes model.

of the δ parameter indicates some weak primacy effects, with the values of the parameter being slightly below zero (−0.13 and −0.09, respectively). Consistent with the results of Experiment 1, the two models yielded similar BIC and AIC values for the all-symptoms presentation condition when information was conveyed numerically (Table 3). For subjects who based their reasoning on verbal information, the temporal Bayes model yielded somewhat lower BIC and AIC values than the standard Bayes model, indicating that in the present study the temporal Bayes model provided a better fit to the data when reasoning was based on verbal terms.

4.2.4. Model-based clustering

We used the same approach as in Experiment 1 to identify homogeneous clusters of subjects. The results shown in Table 4 indicate that only a few subjects were best accounted for by a small positive δ value (i.e., recency), whereas most subjects were assigned to either the standard Bayes (δ = 0) or the primacy (δ < 0) cluster. Note, however, that the size of the δ parameter in the primacy cluster is close to zero (−0.21 and −0.20 in the verbal and numerical condition, respectively), which essentially means that the tendency to overweight earlier evidence was not very pronounced, even for subjects whose judgments were best accounted for by the temporal Bayes model with a δ value slightly below zero.

5. Experiment 3

The results of Experiments 1 and 2 show a remarkably close correspondence between the numerical and verbal condition and also a high accuracy with respect to the posterior probabilities derived from the standard Bayes model. In Experiment 3, we increased the complexity of the task by using unequal prior probabilities for the two cause events. Whereas in the previous experiments both causes were equally likely, $P(Y) = P(X) = 0.5$, in Experiment 3, one cause had a prior probability of 0.33 and the other a prior probability of 0.67. This base rate information was conveyed either numerically or verbally. For the verbal reasoning condition, we used the terms “sometimes” and “frequently,” which were given ratings of 33.13 and

Table 6
Numerical and verbal information provided in Experiment 3.

Probabilities	Numerical condition	Verbal condition
$P(X)$	67%	Frequently
$P(Y)$	33%	Sometimes
$P(A\|X)$	88%	Almost always
$P(A\|Y)$	8%	Almost never
$P(B\|X)$	70%	Often
$P(B\|Y)$	29%	Occasionally
$P(C\|X)$	29%	Occasionally
$P(C\|Y)$	70%	Often
$P(D\|X)$	8%	Almost never
$P(D\|Y)$	88%	Almost always

Note. $P(X)$ and $P(Y)$ are the prior probabilities (base rates) of the cause events, and the conditional probabilities are the individual likelihoods of the four effects (symptoms A, B, C, and D) given each of the two causes (chemicals X and Y). Mapping of words to numbers based on Bocklisch et al. (2012; see Table 1).

Table 7
Two example test trials with three sequentially presented symptoms in Experiment 3. The posterior probabilities derived from the standard Bayes model vary depending on the prior probability of the cause events.

Probability	Low base rate condition $P(Y) = 0.33$		High base rate condition $P(Y) = 0.67$	
	A–D–C	*D–A–B*	*A–D–C*	*D–A–B*
$P(Y)$	0.33	0.33	0.67	0.67
$P(Y\|S_1)$	0.04	0.84	0.16	0.96
$P(Y\|S_1, S_2)$	0.33	0.33	0.67	0.67
$P(Y\|S_1, S_2, S_3)$	0.54	0.17	0.83	0.46

66.11 (out of 100) in the study by Bocklisch et al. (2012; see Table 1). For the numerical condition we used values of 33% and 67%, respectively.

The primary goal was to investigate if subjects considered base rate information in their diagnostic judgments when reasoning based on verbal information. As before, information on the likelihoods (relations between causes and effects) was conveyed either numerically or verbally. We also used a different combination of verbal terms and numerical estimates for the likelihoods from that in the previous studies, thereby further broadening the investigated situations.

5.1. Method

5.1.1. Subjects and design

One hundred sixty-three students from the University of Göttingen participated (112 women, 51 men, $M_{\text{age}} = 23.4$ years); they were paid €8 or received course credit. Subjects were randomly assigned to one of the two (numerical vs. verbal) format conditions. Within each condition we counterbalanced the base rate of the causes between subjects. Mean completion time was 30 min.

5.1.2. Materials and procedure

We used the same cover story, materials, and procedure as in Experiments 1 and 2. The key difference was that in the instructions, subjects were informed that accidents with the two chemicals (cause events) occur with varying frequencies. In the numerical condition subjects received the following information: “Since more workers come in contact with Altexon [Zyroxan], in two thirds of all cases (67%) Altexon [Zyroxan] is the cause of the symptoms; only in one third of the cases (33%) is Zyroxan [Altexon] the cause.” In the verbal condition, subjects were told: “Since more workers come in contact with Altexon [Zyroxan], frequently Altexon [Zyroxan] is the cause of the symptoms; only sometimes is Zyroxan [Altexon] the cause.” Thus, in the verbal condition subjects never received any numerical information on the causes’ base rates.

After reading the instructions, subjects were taught the strength of the individual relations (Table 6) as in the previous studies. As before, subjects had to pass a short instruction test, including a question regarding the base rates of the causes. In the diagnostic reasoning phase, 24 different three-symptom sequences were presented. Table 7 illustrates the predictions for two example trials (symptom sequences *A–D–C* and *D–A–B*), showing how the diagnostic probabilities vary as a function of the causes’ prior probabilities.

In the diagnostic reasoning phase, we additionally requested an initial judgment prior to presenting any symptoms; that is, in each trial, subjects had to judge the probability of a given patient having been in contact with either of the two substances before presenting any symptoms. This prior judgment should reflect the unequal base rates of the causes prior to obtaining any evidence. If subjects ignored the instructed base rates, they should give ratings of around 50 (on the

11-point rating scale of 0–100, corresponding to a probability of 0.5). By contrast, if they appreciated base rates, their initial judgments should reflect the diverging prior probability of the two causes. In total, each subject made 4 × 24 = 96 judgments.

5.2. Results

5.2.1. Learning criterion

We applied the same learning criterion as before, based on testing subjects' knowledge of the strength of the cause–effect relations (likelihoods) after the diagnostic reasoning phase. Forty-four of 163 students were excluded from the analyses (27%) because they could not correctly reproduce at least seven of the eight relations, leaving $N = 119$ valid subjects.

5.2.2. Overall analyses

In total, we obtained 119 (subjects) × 24 (symptom sequences) × 4 (sequential judgments) = 11,424 judgments in the diagnostic reasoning phase. We first analyzed subjects' initial judgments before they were presented with any symptoms. Mean judgments in the numerical condition corresponded closely to the instructed base rates of 33% and 67%: In the low base rate condition, the mean judgment was 0.35 (95% CI [0.31, 0.40]); in the high base rate conditions the mean judgment was 0.65 (95% CI [0.60, 0.69]).

The more interesting analysis concerns the linguistic condition, in which subjects were told only that accidents with the two cause events (chemicals) occur "sometimes" and "frequently." In the study of Bocklisch et al. (2012), the two terms received mean judgments of 33.1 and 66.1 (Table 1). In our study, subjects' mean judgments were 0.37 (95% CI [0.31, 0.43]) and 0.66 (95% CI [0.60, 0.71]), respectively. These findings show a very high consistency between the numerical and verbal condition, and a high consistency in the understanding of the verbal terms between our study and the results obtained by Bocklisch et al. (Note that these analyses show that subjects correctly reported base rates but not necessarily that they utilized them in their diagnostic judgments; we address this issue through additional analyses below.)

For the analyses of the diagnostic judgments, we recoded corresponding trials with varying base rates within each format condition, such that in each trial $P(Y) = 0.33$ and $P(X) = 0.67$. For instance, if $P(Y) = 0.33$ the posterior probabilities entailed by sequence *A–D–C* are complementary to the posterior probabilities entailed by the sequence *D–A–B* when $P(Y) = 0.67$ [e.g., if $P(Y) = 0.33$, then $P(Y|A) = 1 - P(Y|D)$ if $P(Y) = 0.67$; see Table 7].[5] This results in 24 (trial sequences) × 4 (judgments) = 96 mean judgments in each format condition (based on $n = 60$ in the verbal condition and $n = 59$ in the numerical condition). Fig. 8 shows the mean diagnostic judgments for the 24 test trials, including the base rate judgments. As in the previous studies, the results reveal a high consistency between the verbal and numerical condition, and subjects' judgments quite accurately tracked the posterior probabilities derived from the standard Bayes model.

For the subsequent analyses, we excluded the prior judgments, focusing on the diagnostic judgments after symptom information was obtained, yielding 3 (sequential judgments) × 24 (trials) = 72 mean judgments in each condition. The correlation between subjects' diagnostic judgments in the numerical and verbal condition across all diagnostic judgments was $r = 0.992$, $RMSE = 0.042$ (Fig. 6c, left column). The correlation between subjects' judgments in the numerical condition and the predictions of the standard Bayes model was $r = 0.979$, $RMSE = 0.125$; the correlation for the verbal condition was $r = 0.986$, $RMSE = 0.098$ (Fig. 6c, middle and right column).

To investigate people's use of base rate information, we fitted the causes' prior probabilities, using the MSE between the predictions of the standard Bayes model and the mean human judgments. In the numerical condition, the best fitting prior was $P(Y) = 0.35$; in the verbal condition the best fitting prior was 0.37. Both values are close to the true prior of 0.33, and even closer to subjects' mean estimates in the two conditions, which were 0.355 and 0.354 in the numerical and verbal condition, respectively. These results suggest that subjects appropriately utilized base rate information.

5.2.3. Model comparison

Table 3 shows the results for the standard Bayes and the temporal Bayes model, with the δ parameter of the temporal Bayes model fitted to the mean human judgments. The best fitting δ values of the temporal Bayes model (−0.07 in the verbal condition and −0.12 in the numerical condition) are similar to those obtained in Experiment 2, indicating weak primacy effects (i.e., more weight on earlier symptoms). Both models yielded similar BIC and AIC values, consistent with the finding that the δ parameter of the temporal Bayes model was close to zero.

As an additional check for people's use of base rate information, we fitted the δ parameter of the temporal Bayes model and the subjective causes' prior probabilities simultaneously. In the numerical condition, the best fitting prior $P(Y)$ was 0.35 and $\delta = -0.12$; in the verbal condition the best fitting prior was 0.37 and $\delta = -0.08$. Thus, the best fitting priors were very close to the true probability of 0.33, and the size of the δ parameters was unaffected by the base rate estimation (i.e., the two parameters are not interacting). Overall, the results show that subjects adequately utilized base rate information in their

[5] To illustrate, consider Table 7. In the low base rate condition, the probabilities for the sequence *A–D–C* are $P(Y) = 0.33$, $P(Y|A) = 0.04$, $P(Y|A, D) = 0.33$, and $P(Y|A, D, C) = 0.54$. Now consider the sequence *D–A–B* in the high base rate condition, which entails the following probabilities: $P(Y) = 0.67$, $P(Y|A) = 0.96$, $P(Y|A, D) = 0.67$, and $P(Y|A, D, C) = 0.46$. At each step, these probabilities are complementary, e.g., $P(Y|A) = 1 - P(Y|D) = 0.04$. We therefore recoded the corresponding trials in each format condition.

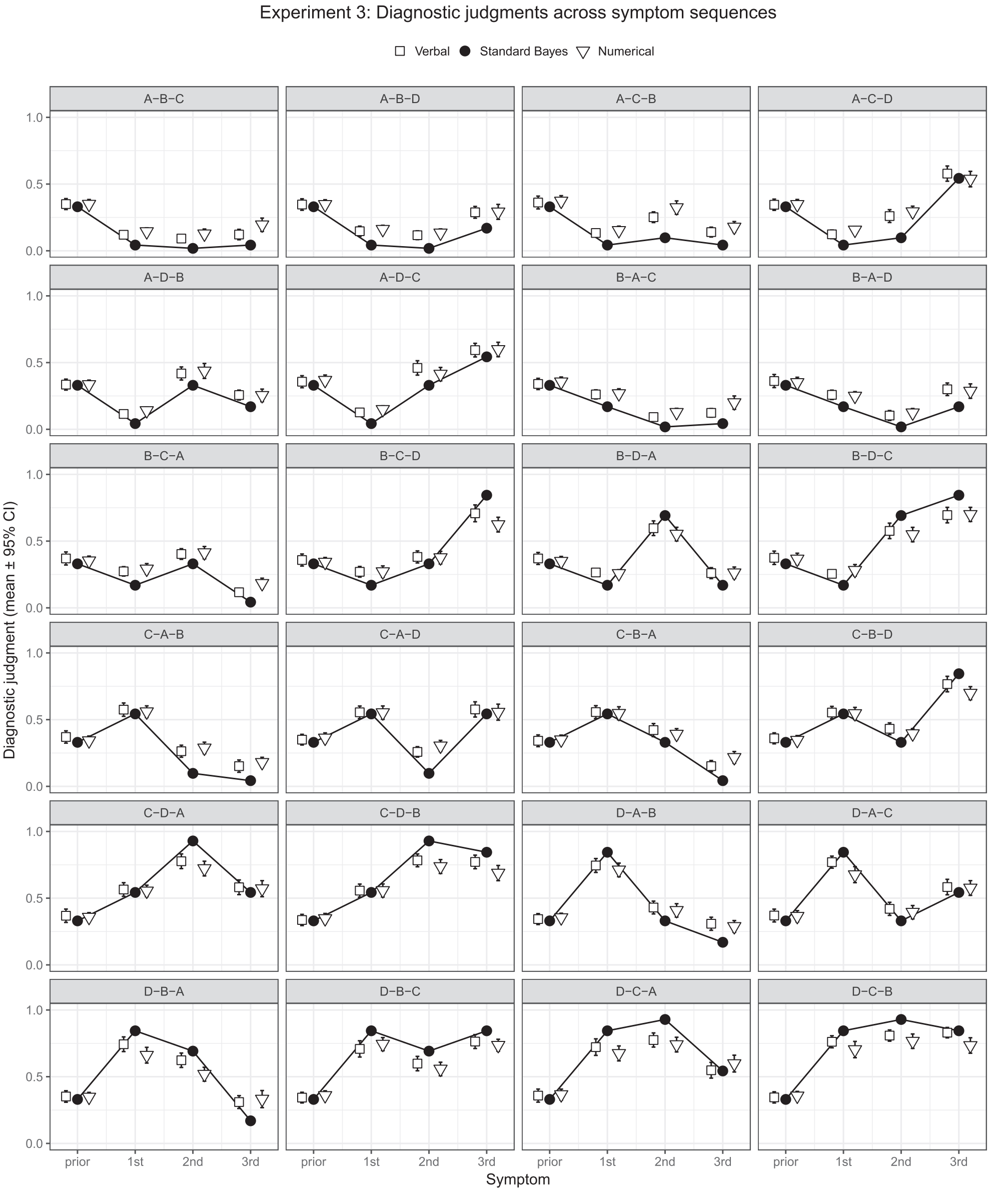


Fig. 8. Mean diagnostic judgments (and 95% confidence intervals) in Experiment 3 and predictions of the standard Bayes model, with the prior probability $P(Y) = 0.33$.

diagnostic inferences, irrespective of whether information on base rate and likelihoods was conveyed numerically or verbally.

5.2.4. Model-based clustering

We used the same cluster technique as before to explore interindividual differences in the temporal weighting of information. Since the previous analyses indicate that people utilized base rate information, we used the same procedure as in Experiments 1 and 2, using the true priors for generating model predictions and defining the three initial clusters by different values of the δ parameter. Consistent with the all-symptoms presentation condition of Experiment 1 and the results of Experiment 2, most subjects were assigned to either the standard Bayes ($\delta = 0$) or the primacy cluster ($\delta < 0$). Overweighting of earlier evidence for subjects assigned to the primacy cluster was not very pronounced, though, as indicated by δ values close to zero (−0.21 and −0.29). In each condition, a minority of subjects was assigned to the recency cluster ($\delta > 0$). In the verbal condition, the tendency to overweight earlier evidence was not very pronounced ($\delta = 0.27$), but the four subjects assigned to the recency cluster in the numerical condition showed a strong tendency to overweight more recent information ($\delta = 2.55$).

6. General discussion

Although verbal terms such as "infrequently," "occasionally," and "sometimes" are rather vague and imprecise, they are commonly used in many real-world situations. In contrast, researchers interested in human probabilistic thinking and judgment under uncertainty usually provide their subjects with precise numerical information or sample data in order to compare their behavior and inferences to the predictions of computational models, which typically also require numerical input.

A key motivation underlying the present work was to investigate subjects' reasoning in situations that more closely resemble real-world situations, in which inferences must often be drawn in the absence of precise and reliable quantitative information. Using a sequential diagnostic reasoning task, we observed that people's inferences were quite consistent when information on cause–effect relations was conveyed through linguistic terms. This by itself is an interesting finding, as it suggests that the verbal expressions were interpreted and utilized in a similar manner across subjects. Even more interesting is that performance was almost indiscernible from performance in a condition in which subjects were provided with matched numerical information. The fact that we took the numerical equivalents from a different study (Bocklisch et al., 2012) supports research showing that the interpretation of linguistic uncertainty terms is relatively stable across populations and contexts (Mosteller & Youtz, 1990; Simpson, 1963; but see Section 6.3 below). This is a promising finding for applying computational models of cognition to verbal reasoning tasks. It is particularly interesting for Bayesian modeling, as this approach is not restricted to numerical point estimates (e.g., mean of an elicited frequency distributions) but can also operate on full distributions (see below).

A theoretical contribution of this paper is the temporal Bayes model, which provides an account for modeling order effects (Bergus et al., 1998; Hogarth & Einhorn, 1992; Rebitschek et al., 2015) within a probabilistic framework. Primacy and recency effects are often considered problematic for Bayesian models of cognition, because the temporal order in which evidence is obtained does not matter for standard Bayesian inference (i.e., the standard Bayes model). Steyvers et al. (2003) presented a Bayesian model that allowed for modeling recency effects (e.g., due to forgetting); our model extends their research by being able to account for both primacy and recency effects. We used the temporal Bayes model to investigate the temporal weighting of evidence in sequential diagnostic reasoning. The model enabled us to quantify the amount of weighting, on both the aggregate and a subgroup level and as a function of task condition (i.e., symptom presentation in Experiment 1). Model-based cluster analyses revealed that most subjects were best accounted for by the standard Bayes model, but in all experiments there were subjects who were better accounted for by a tendency to overweight earlier evidence (primacy) or overweight more recent evidence (recency). Future research should also explore alternative ways of modeling the temporal weighting of information in sequential reasoning. Our model uses an exponential weighting function, but other functions could be used, too (e.g., power functions; Wixted & Ebbesen, 1997). Of course, there are also other ways of modeling order effects within a probabilistic framework, such as using particle filters for modeling primacy effects (Sanborn, Griffiths, & Navarro, 2010). The more general point is that such effects can—and should be—considered when building probabilistic models of cognition, to build bridges between different levels of analysis (Griffiths, Vul, & Sanborn, 2012).

6.1. Linear models of diagnostic judgment

In all three experiments people's diagnostic judgments closely corresponded to the posterior probabilities derived from the Bayesian models, irrespective of whether the relevant information (likelihoods and priors) was conveyed numerically or verbally. Our Bayesian models provide a computational-level explanation of behavior with respect to the reasoner's goals and the probabilistic structure of the environment (Anderson, 1990; Chater & Oaksford, 1999; Chater & Oaksford, 2008; Marr, 1982; for critical reviews, see Brighton & Gigerenzer, 2012; Jones & Love, 2011). These models aim to capture empirical regularities in human behavior, without characterizing the underlying cognitive mechanisms.

Can alternative strategies approximate the predictions of the standard Bayes model and account for people's diagnostic judgments? To address this question, we considered one alternative class of models, namely weighted-additive (WADD) approaches (Brehmer, 1994; Dawes & Corrigan, 1974; Juslin, Nilsson, & Winman, 2009). From this view, the cause event is inferred using an average (i.e., linear) combination of symptom weights ω_t:

$$P(Y|S) = \frac{1}{T}\sum_{t=1}^{T}\omega_t(Y) \tag{7}$$

where $\omega_t(Y)$ denotes the symptom weights of cause Y, t is the current symptom, and T is the total number of symptoms observed so far. The weights are derived from the information people are provided with, that is, likelihoods and priors. We investigated three kinds of linear models that make different assumptions regarding the decision weights. For Experiments 1 and 2, in which the two candidate causes were equiprobable a priori, we considered only the symptoms to make a diagnostic inference (i.e., ignored the causes' prior probabilities). For Experiment 3, in which the causes were not equiprobable a priori, we assumed that this information was utilized by assigning a decision weight ω to them, too.

The simplest model, *tallying*, merely counts symptoms (cf. Dawes, 1979). For instance, in our studies, symptoms *A* and *B* were more likely to be generated by *X*, whereas *C* and *D* were more likely to be generated by *Y* (with the exact likelihoods varying across the three studies). Given a set of symptoms, one simply tallies the evidence. For instance, given the symptom sequence *A*–*C*–*D*, two of the three symptoms provide evidence for *Y* (because *C* and *D* are more likely to be generated by *Y* than by *X*); accordingly, the resulting estimate would be 2/3. Given the likelihoods of Experiment 1, for instance, this estimate is quite close to the true probability of 0.69. Other diagnostic estimates, however, can deviate quite strongly from a Bayesian account. For instance, after observing the first symptom *A*, the tally model would predict with probability 0 that *Y* is the cause (whereas the true probability in Experiment 1 is 0.18), and after the second symptom *C* the model would predict that both causes are equally likely (whereas the true probability is 0.33, given the likelihoods of Experiment 1).

Formally, the tally model considers all symptom likelihoods for which $P(S_t|Y) \neq P(S_t|X)$ and converts them into binary decisions weights:

$$\omega_t(Y) = \begin{cases} 1 & \text{if } P(S_t|Y) > P(S_t|X) \\ 0 & \text{if } P(S_t|Y) < P(S_t|X) \end{cases} \tag{8}$$

and analogously for cause *X*. Thus, if a symptom is more likely to be generated by *Y*, it receives a 1 for *Y* and a 0 for *X*, and vice versa. If the symptom is equally likely under both hypotheses it is assumed to be ignored (which was not the case in any of the present studies). For Experiments 1 and 2 the causes' prior probability was ignored, since $P(X) = P(Y)$. For Experiment 3 we assumed that the prior was considered by also assigning it a binary weight. Consider the condition in which $P(Y) < P(X)$ and assume that symptom *A* is observed, which is more likely to be generated by *X*. The tally model would then predict $P(Y|A) = 1/2$, because the prior speaks for *Y* but the symptom supports cause *X*.

The second linear model we considered assumes that the decision weights ω_t reflect the strength of the cause–effect relations; we therefore call it *likelihood WADD*. This model sums over the likelihoods and normalizes the result by dividing it by the number of presented symptoms. To ensure that the output of the model is valid, that is, in the range [0,1] and $P(Y|S) + P(X|S) = 1$, the likelihoods for each symptom are normalized prior to summing over them:

$$\omega_t(Y) = \frac{P(S_t|Y)}{P(S_t|Y) + P(S_t|X)} \tag{9}$$

and analogously for cause *X*.[6] Note that for the first symptom this model's prediction is identical to the standard Bayes model if $P(X) = P(Y) = 0.5$; also note that this normalization preserves the likelihood ratio for each symptom. For Experiments 1 and 2 with equal priors only the symptom likelihoods are considered. For Experiment 3 with unequal priors, the linear combination (Eq. (7)) includes the prior probability of the causes. [Note that the normalization does not affect the priors, as $P(X) + P(Y) = 1$.]

Finally, we examined the predictions of an "optimal" WADD model, by fitting the weights (likelihoods) to the mean human judgments, using MSE minimization as the criterion. This model is not meant to be a plausible psychological account but serves as a theoretical benchmark since it provides the best fit given the functional form of the model (linear combination) and the data. We used Monte-Carlo simulations with $m = 10^6$ independent samples for each symptom likelihood, drawn from a uniform Beta(1,1) distribution. (For Experiment 3, the symptom likelihoods were randomly generated and the true priors were used, since the empirical results show that people accurately reported the priors and considered them in their diagnostic judgments.)

Table 8 shows the fit of the different linear models for Experiments 1–3, separately for each condition. All models achieved a respectable fit, but none of them could match the Bayesian models, in terms of both correlation and RMSE. These results speak against the idea that our subjects used a linear-additive strategy to make judgments. Note that the likelihood WADD model achieved a quite good fit, consistent with research demonstrating that linear additive strategies can approximate probabilistic inferences (Juslin et al., 2009). However, both in terms of correlation and RMSE it failed to outperform the standard Bayes model (except for the RMSE in the numerical condition in Experiment 3). Also note that even the "optimal" WADD model failed to outperform the standard Bayes account (again with the exception of the RMSE in the numerical condition of Experiment 3), indicating that the functional form of the linear-additive models substantially limits their capacity to account for the human data.

We also examined whether Hogarth and Einhorn's (1992) belief-adjustment model could account for the data. The analyses show that the model was not competitive with the Bayesian models; therefore we do not discuss it here in detail (see Appendix B for details and model fits).

6.2. Modeling the vagueness of verbal terms through probability distributions

Throughout this paper, we have used numerical point estimates (i.e., means of the elicited frequency distributions) of the verbal terms used to derive model predictions. An alternative approach is to use full probability distributions to represent the uncertainty (or vagueness) associated with different verbal terms. This is especially important in light of Wallsten and Budescu's (1995) congruence principle, according to which people aim to match the perceived uncertainty with the choice of communication format. In this view, the vagueness of verbal expressions can be a beneficial feature, as it helps communicate and preserve uncertainty. Erev, Wallsten, and Neal (1991) posited that the vagueness of natural language

[6] An alternative method of bounding the model's prediction to the interval [0,1] would be to use a sigmoidal function, like in logistic regression, and to fit it to subjects' responses. However, our goal was to keep the considered linear models as simple as possible, because the rationale was to test whether simple strategies would provide a better account for the human data than the more complex Bayesian models.

Table 8
Fits of the linear models in Experiments 1–3.

Format	Symptom presentation	Tally		Likelihood WADD		Optimal WADD	
		r	*RMSE*	*r*	*RMSE*	*r*	*RMSE*
Experiment 1 ($N = 112$)							
Verbal	All	0.861	0.167	0.895	0.099	0.899	0.093
	Single	0.817	0.179	0.844	0.161	0.856	0.141
Numerical	All	0.864	0.163	0.889	0.107	0.898	0.097
	Single	0.848	0.169	0.872	0.107	0.881	0.101
Experiment 2 ($N = 61$)							
Verbal	All	0.866	0.172	0.904	0.102	0.911	0.090
Numerical	All	0.883	0.161	0.870	0.117	0.898	0.101
Experiment 3 ($N = 119$)							
Verbal	All	0.867	0.162	0.951	0.120	0.940	0.100
Numerical	All	0.892	0.142	0.959	0.090	0.954	0.070

Note. Predictions for the "optimal" weighted-additive (WADD) model were derived by Monte-Carlo simulations with 10^6 independent samples from a uniform distribution, using the mean-squared error as fitting criterion. r = Pearson correlation, RMSE = Root-mean-squared error.

can also have beneficial effects in social decision-making situations. If there is variability in how people interpret verbal expressions of uncertainty, this can induce heterogeneity in behavior, even if all members of a group are self-interested and pursue the same goal (e.g., maximizing individual utility based on subjective probabilities). Empirically, they found that the use of vague verbal uncertainty expressions increased when the applicable payoff functions were such that heterogeneity in behavior benefited the group. In contrast, when homogeneous behavior was best for the group, communicators preferred to use more precise numerical information.

Previous work has often used membership functions (e.g., Bocklisch et al., 2012; Rapoport et al., 1990; Wallsten, Budescu et al. 1986) based on fuzzy set theory (Zadeh, 1965) to represent within-subject or between-subjects variability of verbal terms. While mathematically well defined, this approach does not neatly integrate with probabilistic models of cognition, since a membership function is not a density function. Probability distributions, by contrast, also allow for representing the vagueness and uncertainty of verbal terms but have the additional advantage that they can be used as input to Bayesian models.

To illustrate, consider Fig. 9. The densities represent the eight different verbal terms used in Experiments 1 and 2, with each term being represented by a Beta distribution. The distributions were fitted to the data from Bocklisch et al. (2012), using the method of moments to derive the shape parameters α and β separately for each term from its sample mean and variance (see Appendix C for details; alternative approaches for fitting probability distributions are possible, of course).

The advantage of representing verbal terms this way is that it preserves the associated uncertainty and variability in mapping words to numbers. In the present case, the between-subjects variability of people's numerical estimates is represented, but the account can also be used to represent within-subject variability (e.g., if repeatedly eliciting numerical estimates from a single subject). Importantly, an advantage of using probability distributions to represent the numerical equivalents of verbal terms is that they can be used as input to probabilistic models. Applied to the diagnostic inference problems considered here, instead of deriving a point estimate for the posterior probability of cause given effect, the full posterior can be derived. Since this account operates on Beta distributions, we here call the model *Beta Bayes*.

Consider the symptom sequence *A*–*D*–*C*. Using Bayes's rule (Eqs. (1) and (2)) to derive the posterior estimate from the likelihoods shown in Fig. 1b and assuming a uniform prior over the causes yields the posterior probabilities 0.18, 0.50, and 0.69 (Table 2). Fig. 10 illustrates the corresponding posterior distributions derived from the Beta distributions representing the verbal terms. The means of these distributions (0.172, 0.500, and 0.689) approximate the point estimates derived from the standard Bayes model, but they additionally represent the associated uncertainty through the variance of the distributions.

For the present set of experiments, the predictions of the standard Bayes model and the Beta Bayes model are quite similar to each other. Accordingly, the fits of the Beta Bayes model are virtually identical to those obtained for the standard Bayes model (Table 9).[7] However, future research should investigate the predictive accuracy of this model for judgment and decision making with verbal terms, as it is capable of explicitly representing the vagueness of natural language uncertainty terms.

[7] Predictions of the Beta Bayes model were derived through Monte-Carlo simulations, with $m = 500{,}000$ independent samples from each Beta distribution representing the symptom likelihoods, with the corresponding shape parameters $\hat{\alpha}$ and $\hat{\beta}$ (see the Appendix C for details). For Experiment 3, in which the causes' priors were unequal, the given point estimates $P(X) = 0.67$ and $P(Y) = 0.33$ were used; likelihoods were sampled from the Beta distributions [since causes are mutually exclusive, the constraint $P(X) + P(Y) = 1$ must hold, which is not the case when sampling the priors from independent Beta distributions. An alternative implementation would be to sample independently from the priors over $P(X)$ and $P(Y)$ and then renormalize].

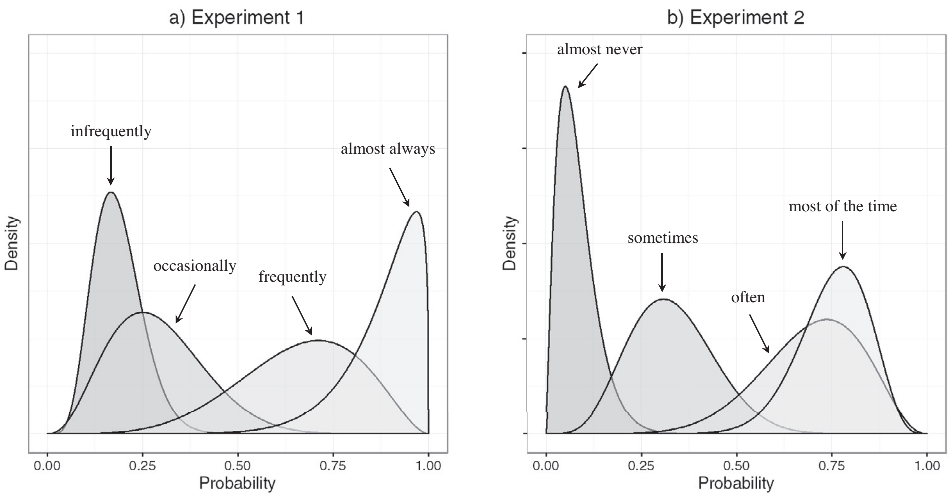


Fig. 9. Beta distributions for the (German) verbal terms used in Experiments 1 and 2. For each distribution, shape parameters α and β were derived using the sample mean and variance from Bocklisch et al. (2012), using the method of moments (see Appendix C for details).

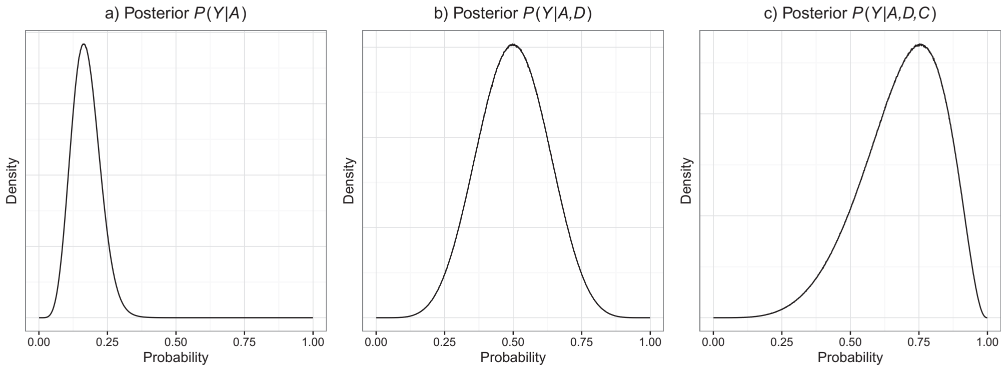


Fig. 10. Posterior probability distributions $P(Y|S)$ for the symptom sequence A–D–C in Experiment 1, with $P(X) = P(Y) = 0.5$. (a) Posterior distribution after observing symptom A. (b) Posterior distribution after observing symptoms A and D. (c) Posterior distribution after observing symptoms A, D, and C.

Table 9
Fits of the Beta Bayes Model in Experiments 1–3.

Format	Symptom presentation	*r*	*RMSE*
Experiment 1 (*N* = 112)			
Verbal	All	0.991	0.029
	Single	0.952	0.096
Numerical	All	0.996	0.022
	Single	0.972	0.052
Experiment 2 (*N* = 61)			
Verbal	All	0.968	0.075
Numerical	All	0.951	0.092
Experiment 3 (*N* = 119)			
Verbal	All	0.989	0.088
Numerical	All	0.983	0.116

Note. Predictions of the Beta Bayes model were derived through Monte-Carlo simulations, with $m = 500{,}000$ independent samples from each Beta distribution representing the symptom likelihoods, with the corresponding shape parameters $\hat{\alpha}$ and $\hat{\beta}$ (see Appendix B for details). r = Pearson correlation, RMSE = Root-mean-squared error.

6.3. Future research and concluding remarks

The present research investigated diagnostic causal inferences based on verbal expressions of uncertainty. Our findings and research methodologies also provide new pathways for investigating empirical phenomena from the literature within a computational modeling framework.

One issue is that the interpretation of verbal expressions can be subject to contextual factors. For instance, the term "frequent" may be interpreted differently in the statements "earthquakes are frequent in California" and "rain is frequent in London" because the overall base rate is lower for the former event than for the latter (Fischer & Jungermann, 1996; Wallsten, Fillenbaum et al., 1986). A related finding is that people's interpretations of verbal terms are influenced by the severity of the outcome (Weber & Hilton, 1990; see Harris, Corner, & Hahn, 2009, for similar findings in situations with an objective frequency basis). Harris and Corner (2011) disentangled base rates and outcome severity and found that people assigned higher numerical values to verbal expressions such as "likely" and "unlikely" when referring to an event with severe negative consequences (e.g., an accident with nuclear waste that would kill thousands of people) than when referring to an event with mild consequences (e.g., an accident with drinking water that would cause a minor traffic delay). Their findings suggest that outcome severity influences the interpretation of verbal terms, independent of base rates.

Addressing these findings in a probabilistic modeling framework is an important issue for future research, both theoretically and empirically. In the present experiments, all events were rather mild symptoms (headache, dizziness, vomiting, fever), so outcome severity did not vary. However, our approach could be used to further investigate how outcome severity influences the understanding and representation of verbal expressions of uncertainty. One way to analyze severity effects would be to use the Beta Bayes model and derive probability distributions of verbal terms in the context of more neutral and more severe events. This could lead to a more comprehensive representation of how people's understanding of verbal terms shifts as a function of outcome severity—for instance, whether only the means of the distributions vary or additionally the amount of uncertainty in the numerical estimates changes (e.g., variance of the distribution). Such data could also inform empirical research. Using the methodology of the current studies, one could elicit numerical estimates for different expressions in the context of more neutral and more severe outcomes (e.g., the chemical *frequently* leads to loss of eyesight vs. the chemical *frequently* leads to headache) and then use these estimates to derive quantitative predictions for probabilistic diagnostic inferences with verbal information. If people assign different numerical estimates as a function of outcome severity, different inferences should result compared to a group in which the same terms are used in combination with less severe outcomes.

A second key issue is that in everyday communication, communicating uncertainty does not rest solely on the use of natural language uncertainty terms; pragmatics and conversational goals also play a role. For instance, in a medical context the word "possibly" could be interpreted as a likelihood statement (i.e., how likely it is to develop a severe medical condition) or it could be perceived as a politeness marker (i.e., as a face-saving device when communicating bad news). Bonnefon and Villejoubert (2006; see also Bonnefon, Feeney, & De Neys, 2011) tested this idea empirically. Participants were asked to provide numerical estimates for the word "possibly" in a statement such as "The doctor tells you, you will possibly suffer from deafness soon." Using membership functions to model participants' numerical estimates, the critical finding was that the membership function for people who considered the statement a face-saving device peaked at a higher probability than for those who considered it to mean merely communicating a likelihood estimate (for related findings with numerical information, see Sirota & Juanchich, 2012). This result shows that the understanding of verbal terms in social situations is also mediated by conversational rules and assumptions about the goals of the communicator.

Pragmatics and conversational rules should not have played a role in the scenario used in the present studies. The more general question of how to incorporate conversational pragmatics into a probabilistic modeling framework is an important issue, though (Harris, Corner, & Hahn, 2014). From a Bayesian perspective, one could model politeness effects by explicitly representing hypotheses about the goals of the communicator, such as whether a statement is used to communicate likelihood information or serves as a face-saving device. Depending on the relative plausibility of these hypotheses, different numerical estimates could be derived via Bayesian model averaging (see Meder et al., 2014, for such an approach in the domain of diagnostic reasoning). Consider a patient who is being told by the doctor that "it is *likely* that you suffer from a serious medical condition." The term may be understood as a likelihood estimate, but the patient may also wonder whether the communicator is downplaying that in fact having the condition is *very likely*. According to the meta-analysis of Mosteller and Youtz (1990), the corresponding numerical probabilities of *likely* and *very likely* are 0.69 and 0.82. Depending on how much weight is assigned to each of these hypotheses, a weighted average of the two estimates could be formed that reflects the reasoner's uncertainty about the goals of the speaker (e.g., a reasoner who is maximally uncertain about the goal of the communicator may assign equal probability to each hypothesis and infer a numerical value of $0.5 \cdot 0.69 + 0.5 \cdot 0.82 = 0.755$ for the likelihood of the event). Such an account would incorporate uncertainty about conversational pragmatics and the goals of the communicator.

In sum, we consider it an important issue to investigate how people understand and utilize natural language expressions of uncertainty, both theoretically and empirically. Whereas the use of verbal terms is ubiquitous in many real-world situations, they do not easily fit with formal models of cognition, which usually require precise numerical information. The present research provides new pathways for modeling reasoning with verbal information, which will help researchers design experiments that more closely correspond to the ways people communicate outside the lab, in order to gain a better understanding of how they reason and decide on the basis of verbal frequency and probability terms.

Funding

This research was supported by grants ME 3717/2-2, MA 6545/1-2, and WA 621/22-1 from the Deutsche Forschungsgemeinschaft as part of the priority program “New Frameworks of Rationality” (SPP 1516).

Acknowledgments

We thank Franziska Bocklisch, Marc Buehner, Ulrike Hahn, Georg Jahn, Felix Rebitschek, and Michael Waldmann for helpful comments, and Anita Todd for editing the manuscript.

Appendix A

In the following we provide numerical examples for computing posterior probabilities for the symptom sequence *A–C–D* in Experiment 1, assuming $P(X) = P(Y) = 0.5$. We first consider the standard Bayes model and then the temporal Bayes model. The individual likelihoods of causes *X* and *Y* for the three effects *A*, *C*, and *D* are as follows (see Fig. 1): $P(A|X) = 0.88$ and $P(A|Y) = 0.19$, $P(C|X) = 0.29$ and $P(C|Y) = 0.66$, and $P(D|X) = 0.19$ and $P(D|Y) = 0.88$. The goal is to infer the posterior probabilities of cause *Y* given the sets of symptoms $S_1 = \{A\}$, $S_2 = \{A, C\}$, and $S_3 = \{A, C, D\}$. Other posterior probabilities can be computed accordingly by plugging in the respective prior probabilities and likelihoods (see Fig. 1 for the likelihoods used in Experiments 1 and 2, which assumed a uniform prior over the causes, $P(X) = P(Y) = 0.5$; see Table 6 for the prior probabilities and individual likelihoods used in Experiment 3).

A.1. Standard Bayes model

We derive posterior probabilities first using the standard form of Bayes’s rule and then using the log odds form of Bayes’s rule. According to Bayes’s rule in its standard form,

$$P(Y|S) = \frac{P(S|Y)\ \cdot\ P(Y)}{P(S|Y)\ \cdot\ P(Y) + P(S|X)\ \cdot\ P(X)} \tag{A1}$$

where $P(S|Y)$ and $P(S|X)$ denote the likelihoods of the observed symptoms given causes *X* and *Y*, respectively, and $P(Y)$ and $P(X)$ denote the prior probabilities of the causes.

The posterior probability of cause *Y* given symptom *A* is

$$P(Y|A) = \frac{P(A|Y) \cdot P(Y)}{P(A|Y)\ \cdot P(Y) + P(A|X)\ \cdot P(X)} = \frac{0.19\ \cdot 0.5}{0.19\ \cdot 0.5 + 0.88\ \cdot 0.5} = 0.1775701 \approx 0.18$$

We next derive the posterior probability of *Y* given symptom sets $S_2 = \{A, C\}$, and $S_3 = \{A, C, D\}$. Assuming that the causal Markov condition (Pearl, 2000) holds, the joint likelihood of a symptom set *S* given cause *Y* can be computed from the product of the individual likelihoods S_t:

$$P(S|Y) = \prod_{t=1}^{T} P(S_t|Y) \tag{A2}$$

Accordingly,

$$\begin{aligned} P(Y|A, C) &= \frac{P(A, C|Y) \cdot P(Y)}{P(A, C|Y) \cdot P(Y) + P(A, C|X) \cdot P(X)} \\ &= \frac{P(A|Y) \cdot P(C|Y) \cdot P(Y)}{P(A|Y) \cdot P(C|Y) \cdot P(Y) + P(A|X) \cdot P(C|X) \cdot P(X)} \\ &= \frac{0.19 \cdot 0.66 \cdot 0.5}{0.19 \cdot 0.66 \cdot 0.5 + 0.88 \cdot 0.29 \cdot 0.5} = 0.3294798 \approx .033 \end{aligned}$$

The posterior probability of cause *Y* given symptoms *A*, *C*, and *D* can be derived analogously:

$$\begin{aligned} P(Y|A, C, D) &= \frac{P(A, C, D|Y) \cdot P(Y)}{P(A, C, D|Y) \cdot P(Y) + P(A, C, D|X) \cdot P(X)} \\ &= \frac{P(A|Y) \cdot P(C|Y) \cdot P(D|Y) \cdot P(Y)}{P(A|Y) \cdot P(C|Y) \cdot P(D|Y) \cdot P(Y) + P(A|X) \cdot P(C|X) \cdot P(D|X) \cdot P(X)} \\ &= \frac{0.19 \cdot 0.66 \cdot 0.88 \cdot 0.5}{0.19 \cdot 0.66 \cdot 0.88 \cdot 0.5 + 0.88 \cdot 0.29 \cdot 0.19 \cdot 0.5} = 0.6947368 \approx 0.69 \end{aligned}$$

A.2. Standard Bayes Model: Log odds form

An alternative way of computing these posteriors is by using the log odds form of Bayes's rule:

$$\varphi = \log\frac{P(Y|S)}{P(X|S)} = \log\frac{P(Y)}{P(X)} + \sum_{t=1}^{T}\log\frac{P(S_t|Y)}{P(S_t|X)} \tag{A3}$$

where the first summand is the prior odds and the second summand is the likelihood odds and T is the total number of symptoms. For symptom sets $S_1 = \{A\}$, $S_2 = \{A, C\}$, and $S_3 = \{A, C, D\}$, this yields

$$\varphi_{S_1=\{A\}} = \log\frac{P(Y|A)}{P(X|A)} = \log\frac{P(Y)}{P(X)} + \log\frac{P(A|Y)}{P(A|X)} = \log\frac{0.5}{0.5} + \log\frac{0.19}{0.88} = -1.532898$$

$$\varphi_{S_2=\{A,C\}} = \log\frac{P(Y|A,D)}{P(X|A,D)} = \log\frac{P(Y)}{P(X)} + \log\frac{P(A|Y)}{P(A|X)} + \log\frac{P(C|Y)}{P(C|X)} = \log\frac{0.5}{0.5} + \log\frac{0.19}{0.88} + \log\frac{0.66}{0.29} = -0.7105389$$

$$\varphi_{S_3=\{A,C,D\}} = \log\frac{P(Y|A,D,C)}{P(X|A,D,C)} = \log\frac{P(Y)}{P(X)} + \log\frac{P(A|Y)}{P(A|X)} + \log\frac{P(C|Y)}{P(C|X)} + \log\frac{P(D|Y)}{P(D|X)} = \log\frac{0.5}{0.5} + \log\frac{0.19}{0.88} + \log\frac{0.66}{0.29} + \log\frac{0.88}{0.19} = 0.8223589$$

where log denotes the natural logarithm. The log posterior odds can be transformed into a conditional probability by an inverse-logit transformation:

$$P(Y|S) = \frac{1}{1 + e^{-\varphi}} \tag{A4}$$

For instance, the posterior probability of cause Y given symptom A is given by

$$P(Y|A) = \frac{1}{1 + e^{-(-1.532898)}} = 0.1775701 \approx 0.18$$

which is identical to the value computed above via the standard form of Bayes's rule. Analogously, the posterior probability of Y given symptom sets $S_2 = \{A, C\}$ and $S_3 = \{A, C, D\}$, respectively, are given by

$$P(Y|A,C) = \frac{1}{1 + e^{-(-0.7105389)}} = 0.3294798 \approx 0.33$$

and

$$P(Y|A,C,D) = \frac{1}{1 + e^{-(-0.8223589)}} = 0.6947368 \approx 0.69$$

which are identical to the values computed above.

A.3. Temporal Bayes model

We next show how to derive the posterior probabilities according to the temporal Bayes model. The log odds form of the standard Bayes model allows for introducing an exponential weighting parameter δ that controls the weighting of symptoms as a function of the temporal order in which they are observed. The general form of the model is

$$\varphi = \log\frac{P(Y)}{P(X)} + \sum_{t=1}^{T} w_\delta(t) \cdot \log\frac{P(S_t|Y)}{P(S_t|X)} \tag{A5}$$

where t is the current symptom and T is the total number of symptoms observed so far. The weighting function $w_\delta(t)$ for a given δ and T is given by

$$w_\delta(t) = \begin{cases} e^{\delta(t-T)} & \text{if } \delta > 0 \\ e^0 = 1 & \text{if } \delta = 0 \\ e^{\delta(t-1)} & \text{if } \delta < 0 \end{cases} \tag{A6}$$

Fig. 2 illustrates the weighting function for different values of δ. For the numerical example, we parameterize the model with $\delta = 1$, in which case more recent evidence is weighted more strongly. We derive posterior probabilities for the same symptom sequence as above [i.e., sequence *A*–*C*–*D* in Experiment 1, assuming $P(X) = P(Y) = 0.5$].

As noted in the main text (Section 2.2), for the first symptom (or if there is only a single symptom) the temporal Bayes model is equivalent to the log odds form of the standard Bayes model, irrespective of the value of δ. For the first symptom $w_\delta(t = 1) = e^0 = 1$. Accordingly,

$$\varphi_{S_1=\{A\}} = \log\frac{P(Y|A)}{P(X|A)} = \log\frac{P(Y)}{P(X)} + w_\delta(t=1) \cdot \log\frac{P(A|Y)}{P(A|X)} = \log\frac{0.5}{0.5} + 1 \cdot \log\frac{0.19}{0.88} = -1.532898$$

which is identical to the log odds form of the standard Bayes model.

For the next judgment, there are two symptoms, *A* and *C*. In this case, the predictions of the temporal Bayes model diverge from those of the standard model, because symptom log odds are weighted differently. According to the weighting function, the two symptom weights for the first ($t = 1$) and second ($t = 2$) symptom are

$$w_\delta(t=1) = e^{\delta(t-T)} = e^{\delta(1-2)} = 0.3678794$$
$$w_\delta(t=2) = e^{\delta(t-T)} = e^{\delta(2-2)} = 1$$

Plugging these values into the temporal Bayes model yields the following:

$$\begin{aligned}\varphi_{S_2=\{A,C\}} &= \log\frac{P(Y|A,C)}{P(X|A,C)} = \log\frac{P(Y)}{P(X)} + w_\delta(t=1)\cdot\log\frac{P(A|Y)}{P(A|X)} + w_\delta(t=2)\cdot\log\frac{P(C|Y)}{P(C|X)}\\ &= \log\frac{0.5}{0.5} + .3678794\cdot\log\frac{0.19}{0.88} + 1\cdot\log\frac{0.66}{0.29} = 0.2584374\end{aligned}$$

Using the inverse-logit transformation (Eq. (A4)), this yields $P(Y|A,C) \approx 0.56$, which is higher than the posterior probability derived from the standard Bayes model, according to which $P(Y|A,C) = 0.33$. This difference arises because the temporal Bayes model with $\delta = 1$ places more weight on the more recent symptom *C*, which is more likely given cause *Y* than given cause *X* (i.e., the log odds of symptom *C* receive a higher weight than the log odds of symptom *A*).

For the full symptom sequence *A*–*C*–*D*, the computation is done analogously. The weights are given by

$$w_\delta(t=1) = e^{\delta(t-T)} = e^{\delta(1-3)} = 0.1353353$$
$$w_\delta(t=2) = e^{\delta(t-T)} = e^{\delta(2-3)} = 0.3678794$$
$$w_\delta(t=3) = e^{\delta(t-T)} = e^{\delta(3-3)} = 1$$

yielding

$$\begin{aligned}\varphi_{S_3=\{A,C,D\}} &= \log\frac{P(Y|A,C,D)}{P(X|A,C,D)}\\ &= \log\frac{P(Y)}{P(X)} + w_\delta(t=1)\cdot\log\frac{P(A|Y)}{P(A|X)} + w_\delta(t=2)\cdot\log\frac{P(C|Y)}{P(C|X)} + w_\delta(t=3)\cdot\log\frac{P(D|Y)}{P(D|X)}\\ &= \log\frac{0.5}{0.5} + 0.1353353\cdot\log\frac{0.19}{0.88} + 0.3678794\cdot\log\frac{0.66}{0.29} + 1\cdot\log\frac{0.88}{0.19} = 1.627972\end{aligned}$$

Using the inverse-logit transformation (Eq. (A4)), this yields $P(Y|A, C, D) \approx 0.86$, which is higher than the posterior probability according to the standard Bayes model, according to which $P(Y|A, C, D) = 0.69$. As with two symptoms, this divergence results from a higher weight on more recent evidence, which favors cause *Y* over cause *X* (because both symptoms *C* and *D* are more likely given cause *Y* than given cause *X*).

Appendix B

The belief-adjustment model (BAM; Hogarth & Einhorn, 1992) can be written as either an adding model, that is, assuming that people encode evidence in an absolute manner, or as an averaging model, that is, assuming that people encode evidence relative to the current belief (for a detailed derivation see Trueblood & Busemeyer, 2011, p. 1539).

Let $S_k \epsilon [0, 1]$ denote the reasoner's belief in the target hypothesis (here: that chemical *Y* caused the symptoms) at time step *k*, let $\alpha\epsilon[0, 1]$ and $\beta\epsilon[0, 1]$ denote the sensitivity to positive and negative evidence, respectively, and let $s(x_k)\epsilon[-1, 1]$ denote the weight of the evidence (here: symptom) presented at step *k*. The adding variant of BAM can then be written as

$$S_k = \begin{cases} S_{k-1} + \alpha\cdot S_{k-1}\cdot s(x_k) & |s(x_k) \leqslant 0 \\ S_{k-1} + \beta\cdot(1-S_{k-1})\cdot s(x_k) & |s(x_k) > 0 \end{cases} \tag{B1}$$

Analogously and assuming $s(x_k)\varepsilon[0, 1]$ as the evidence weights, the averaging variant of BAM is given by

$$S_k = \begin{cases} S_{k-1} + \alpha\cdot S_{k-1}\cdot (s(x_k) - S_{k-1}) & |s(x_k) \leqslant S_{k-1} \\ S_{k-1} + \beta\cdot(1-S_{k-1})\cdot (s(x_k) - S_{k-1}) & |s(x_k) > S_{k-1} \end{cases} \tag{B2}$$

Note that for our experiments both model variants have six free parameters, namely, $s(A)$, $s(B)$, $s(C)$, and $s(D)$, which are the evidence weights of the four different symptoms, and the two sensitivity parameters α and β. We fitted both variants to the data of Experiments 1–3, assuming that the initial belief state S_0 (i.e., the belief prior to observing any symptoms) reflects the cause's prior probability. Accordingly, for Experiments 1 and 2 we assumed the initial belief state S_0 to be 0.5; for Experiment 3 we assumed it to be 0.33 (see main text for details).

The six free parameters were fitted using Monte-Carlo simulations with $m = 10^8$ samples; that is, we drew 10^8 parameter tuples from independent uniform distributions over the parameters' domain of definition and selected the tuple that minimized the mean-squared error for the data (mean human judgments) of each of the experiments' conditions.

Table B1
Fit measures for the belief-adjustment model (adding vs. averaging variant) in Experiments 1–3.

Format	Symptom presentation	BAM adding				BAM averaging			
		r	*RMSE*	AIC	BIC	*r*	*RMSE*	AIC	BIC
Experiment 1 ($N = 112$)									
Verbal	All	0.922	0.084	−77.00	−71.66	0.840	0.118	−64.85	−59.51
	Single	0.971	0.066	−85.88	−80.54	0.937	0.100	−70.77	−65.42
Numerical	All	0.942	0.076	−80.85	−75.50	0.866	0.113	−66.42	−61.08
	Single	0.974	0.049	−96.75	−91.41	0.929	0.080	−78.72	−73.38
Experiment 2 ($N = 61$)									
Verbal	All	0.870	0.114	−144.63	−135.13	0.787	0.149	−124.94	−115.44
Numerical	All	0.881	0.116	−143.06	−133.56	0.785	0.153	−122.93	−113.43
Experiment 3 ($N = 119$)									
Verbal	All	0.913	0.100	−319.96	−306.30	0.807	0.151	−259.97	−246.31
Numerical	All	0.921	0.083	−346.96	−333.30	0.827	0.126	−286.80	−273.14

Note. Predictions for the two variants of the belief-adjustment model (BAM) were derived by Monte-Carlo simulations with 10^8 independent samples from uniform distributions over the six parameters' domain, using the mean-squared error as fitting criterion. *r* = Pearson correlation; RMSE = root-mean-squared error; AIC = Akaike information criterion; BIC = Bayesian information criterion.

The obtained fit measures are shown in Table B1. A comparison with the fits of the standard and temporal Bayes models (Table 3) shows that the BAM model fails to provide an adequate account of the human data. Despite having six free parameters, the BAM model variants were outperformed in most conditions by the standard and temporal Bayes models in terms of the correlation with the judgments and the RMSE (cf. Table 3). Because of the large number of free parameters, the BIC and AIC values (which take into account the number of free parameters) are most informative. The averaging variant of the BAM model was competitive with neither the standard Bayes nor the temporal Bayes model, both of which had a better fit (= lower AIC and BIC values) in each condition of each experiment. The adding variant performed slightly better, but it achieved a better fit than the two Bayesian models only in the numerical condition of Experiment 3 and tied with the standard Bayes model in the single-symptom verbal condition of Experiment 1. Taken together, these findings suggest that the BAM model does not provide a good account of the empirical data.

Appendix C

We fitted Beta distributions to the data from Bocklisch et al. (2012), using the method of moments to derive the shape parameters α and β separately for each verbal term from its sample mean $\bar{x}$ and sample variance s^2. The estimate for shape parameter α, $\hat{\alpha}$, is given by

$$\hat{\alpha} = \bar{x}\left(\frac{\bar{x}(1-\bar{x})}{s^2} - 1\right) \tag{A1}$$

and the estimate for shape parameter β, $\hat{\beta}$, is given by

$$\hat{\beta} = (1-\bar{x})\left(\frac{\bar{x}(1-\bar{x})}{s^2} - 1\right) \tag{A2}$$

Table C1 shows the estimates of the shape parameters for the different frequency expressions (see Fig. 9). For each term, the corresponding Beta distribution has the same mean and variance as the empirical distribution from Bocklisch et al. (2012).

Table C1
Estimates of shape parameters α and β of Beta distributions for numerical estimates of 10 verbal frequency expressions from Bocklisch et al. (2012).

Frequency expression (original German)	Mean $\bar{x}$	Variance s^2	$\hat{\alpha}$	$\hat{\beta}$
Never (nie)	0.014	0.0005	0.359	26.813
Almost never (fast nie)	0.083	0.0025	2.419	26.696
Infrequently (selten)	0.185	0.0040	6.724	29.582
Occasionally (gelegentlich)	0.289	0.0150	3.685	9.058
Sometimes (manchmal)	0.331	0.0120	5.779	11.664
In half of the cases (in der Hälfte der Fälle)	0.501	0.0001	855.649	850.871
Frequently (häufig)	0.661	0.0238	5.560	2.850
Often (oft)	0.697	0.0167	8.137	3.544
Most of the time (meistens)	0.755	0.0082	16.307	5.303
Almost always (fast immer)	0.881	0.0089	9.433	1.273
Always (immer)	0.975	0.0038	5.363	0.140

Note. Words in parentheses denote the corresponding German terms used in Bocklisch et al. (2012) and our studies. Human judgments were given in a frequency format (e.g., "in *X* of 100 cases"); we divided them by 100 to map them onto the interval [0, 1]. Shape parameters $\hat{\alpha}$ and $\hat{\beta}$ were computed using the exact variance estimates, not the rounded values reported in this table.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski (Eds.), *Proceedings of the second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Akademiai Kiado.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Association of Forensic Science Providers (2009). Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice, 49*, 161–164. http://dx.doi.org/10.1016/j.scijus.2009.07.004.

Bergus, G. R., Chapman, G. B., Levy, B. T., Ely, J. W., & Oppliger, R. A. (1998). Clinical diagnosis and the order of information. *Medical Decision Making, 18*, 412–417. http://dx.doi.org/10.1177/0272989X9801800409.

Berry, D. C., Knapp, P. R., & Raynor, T. (2002). Is 15 per cent very common? Informing people about the risks of medication side effects. *International Journal of Pharmacy Practice, 10*, 145–151. http://dx.doi.org/10.1111/j.2042-7174.2002.tb00602.x.

Bocklisch, F., Bocklisch, S. F., & Krems, J. F. (2012). Sometimes, often, and always: Exploring the vague meanings of frequency expressions. *Behavior Research Methods, 44*, 144–157. http://dx.doi.org/10.3758/s13428-011-0130-8.

Bonnefon, J.-F., Feeney, A., & De Neys, W. (2011). The risk of polite misunderstandings. *Current Directions in Psychological Science, 20*, 321–324. http://dx.doi.org/10.1177/0963721411418472.

Bonnefon, J.-F., & Villejoubert, G. (2006). Tactful or doubtful? Expectations of politeness explain the severity bias in the interpretation of probability phrases. *Psychological Science, 17*, 747–751. http://dx.doi.org/10.1111/j.1467-9280.2006.01776.x.

Brehmer, B. (1994). The psychology of linear judgement models. *Acta Psychologica, 87*, 137–154. http://dx.doi.org/10.1016/0001-6918(94)90048-5.

Brighton, H., & Gigerenzer, G. (2012). Are rational actor models "rational" outside small worlds? In S. Okasha, & K. Binmore (Eds.), *Evolution and rationality: Decisions, co-operation, and strategic behaviour* (pp. 84–109). http://dx.doi.org/10.1017/CBO9780511792601.006.

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes, 41*, 390–404. http://dx.doi.org/10.1016/0749-5978(88)90036-2.

Budescu, D. V., Broomell, S., & Por, H.-H. (2009). Improving communication of uncertainty in the reports of the intergovernmental panel on climate change. *Psychological Science, 20*, 299–308. http://dx.doi.org/10.1111/j.1467-9280.2009.02284.x.

Budescu, D. V., Por, H.-H., Broomell, S. B., & Smithson, M. (2014). The interpretation of IPCC probabilistic statements around the world. *Nature Climate Change, 4*, 508–512. http://dx.doi.org/10.1038/nclimate2194.

Budescu, D. V., & Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes, 36*, 391–405. http://dx.doi.org/10.1016/0749-5978(85)90007-X.

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 281–294. http://dx.doi.org/10.1037/0096-1523.14.2.281.

Burnham, K. P., & Anderson, D. R. (1998). *Model selection and multimodel inference: A practical information-theoretic approach*. New York, NY: Springer.

Chapman, G. B., Bergus, G. R., & Elstein, A. S. (1996). Order of information affects clinical judgment. *Journal of Behavioral Decision Making, 9*, 201–211. http://dx.doi.org/10.1002/(SICI)1099-0771(199609)9:3<201::AID-BDM229>3.0.CO;2-J.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences, 3*, 57–65. http://dx.doi.org/10.1016/S1364-6613(98)01273-X.

Chater, N., & Oaksford, M. (Eds.). (2008). *The probabilistic mind*. New York, NY: Oxford University Press.

Clark, D. A. (1990). Verbal uncertainty expressions: A critical review of two decades of research. *Current Psychology, 9*, 203–235. http://dx.doi.org/10.1007/BF02686861.

Cliff, N. (1959). Adverbs as multipliers. *Psychological Review, 66*, 27–44. http://dx.doi.org/10.1037/h0045660.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571–582. http://dx.doi.org/10.1037/0003-066X.34.7.571.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95–106. http://dx.doi.org/10.1037/h0037613.

de Keijser, J., & Elffers, H. (2012). Understanding of forensic expert reports by judges, defense lawyers and forensic professionals. *Psychology, Crime & Law, 18*, 191–207. http://dx.doi.org/10.1080/10683161003736744.

Dhami, M. K., & Wallsten, T. S. (2005). Interpersonal comparison of subjective probabilities: Toward translating linguistic probabilities. *Memory & Cognition, 33*, 1057–1068. http://dx.doi.org/10.3758/BF03193213.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning, 29*, 103–130. http://dx.doi.org/10.1023/A:1007413511361.

Du, N., Budescu, D. V., Shelly, M. K., & Omer, T. C. (2011). The appeal of vague financial forecasts. *Organizational Behavior and Human Decision Processes, 114*, 179–189. http://dx.doi.org/10.1016/j.obhdp.2010.10.005.

Erev, I., & Cohen, B. L. (1990). Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. *Organizational Behavior and Human Decision Processes, 45*, 1–18. http://dx.doi.org/10.1016/0749-5978(90)90002-Q.

Erev, I., Wallsten, T. S., & Neal, M. M. (1991). Vagueness, ambiguity, and the cost of mutual understanding. *Psychological Science, 2*, 321–324. http://dx.doi.org/10.1111/j.1467-9280.1991.tb00159.x.

European Network of Forensic Science Institutes (2015). ENFSI Guideline for evaluative reporting in forensic science Retrieved from <http://www.enfsi.eu/news/enfsi-guideline-evaluative-reporting-forensic-science>.

European Union (2009). A guideline on summary of product characteristics Retrieved from <http://ec.europa.eu/health/files/eudralex/vol-2/c/smpc_guideline_rev2_en.pdf>.

Fenton, N., Neil, M., & Lagnado, D. A. (2012). A general structure for legal arguments about evidence using Bayesian networks. *Cognitive Science, 37*, 61–102. http://dx.doi.org/10.1111/cogs.12004.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science, 21*, 329–336. http://dx.doi.org/10.1177/0956797610361430.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General, 140*, 168–185. http://dx.doi.org/10.1037/a0022100.

Fischer, K., & Jungermann, H. (1996). Rarely occurring headaches and rarely occurring blindness: Is rarely= rarely? The meaning of verbal frequentistic labels in specific medical contexts. *Journal of Behavioral Decision Making, 9*, 153–172. http://dx.doi.org/10.1002/(SICI)1099-0771(199609)9:3<153::AID-BDM222>3.0.CO;2-W.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684–704. http://dx.doi.org/10.1037/0033-295X.102.4.684.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science, 21*, 263–268. http://dx.doi.org/10.1177/0963721412447619.

Hamm, R. M. (1991). Selection of verbal probabilities: A solution for some problems of verbal probability expression. *Organizational Behavior and Human Decision Processes, 48*, 193–223. http://dx.doi.org/10.1016/0749-5978(91)90012-I.

Harris, A. J. L., & Corner, A. (2011). Communicating environmental risks: Clarifying the severity effect in interpretations of verbal probability expressions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 1571–1578. http://dx.doi.org/10.1037/a0024195.

Harris, A. J. L., Corner, A., & Hahn, U. (2009). Estimating the probability of negative events. *Cognition, 110*, 51–64. http://dx.doi.org/10.1016/j.cognition.2008.10.006.

Harris, A. J. L., Corner, A., & Hahn, U. (2014). James is polite and punctual (and useless): A Bayesian formalisation of faint praise. *Thinking & Reasoning, 19*, 414–429. http://dx.doi.org/10.1080/13546783.2013.801367.

Harris, A. J. L., Corner, A., Xu, J., & Du, X. (2013). Lost in translation? Interpretations of the probability phrases used by the Intergovernmental Panel on Climate Change in China and the UK. *Climatic Change, 121*, 415–425. http://dx.doi.org/10.1007/s10584-013-0975-1.

Hayes, B. K., Hawkins, G. E., Newell, B. R., Pasqualino, M., & Rehder, B. (2014). The role of causal models in multiple judgments under uncertainty. *Cognition, 133*, 611–620. http://dx.doi.org/10.1016/j.cognition.2014.08.011.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24*, 1–55. http://dx.doi.org/10.1016/0010-0285(92)90002-J.

Intergovernmental Panel on Climate Change (2007). Contribution of working groups I, II and III to the fourth assessment report of the Intergovernmental Panel on Climate Change. Retrieved from <http://www.ipcc.ch/publications_and_data/ar4/syr/en/contents.html>.

Jarecki, J., Meder, B., & Nelson, J. D. (2013). The assumption of class-conditional independence in category learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (pp. 2650–2655). Austin, TX: Cognitive Science Society.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences, 34*, 169–188. http://dx.doi.org/10.1017/S0140525X10003134.

Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review, 116*, 856–874. http://dx.doi.org/10.1037/a0016979.

Kerstholt, J. H., & Jackson, J. L. (1998). Judicial decision making: Order of evidence presentation and availability of background information. *Applied Cognitive Psychology, 12*, 445–454. http://dx.doi.org/10.1002/(SICI)1099-0720(199810)12:5<445::AID-ACP518>3.0.CO;2-8.

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

Lichtenstein, S., & Newman, J. R. (1967). Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychonomic Science, 9*, 563–564. http://dx.doi.org/10.3758/BF03327890.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: Freeman.

Martire, K. A., Kemp, R. I., Sayle, M., & Newell, B. R. (2014). On the interpretation of likelihood ratios in forensic science evidence: Presentation formats and the weak evidence effect. *Forensic Science International, 240*, 61–68. http://dx.doi.org/10.1016/j.forsciint.2014.04.005.

Mayrhofer, R., & Waldmann, M. R. (2016). Sufficiency and necessity assumptions in causal structure induction. *Cognitive Science, 40*, 2137–2150. http://dx.doi.org/10.1111/cogs.12318.

Meder, B., & Gigerenzer, G. (2014). Statistical thinking: No one left behind. In E. J. Chernoff & B. Sriraman (Eds.), *Probabilistic thinking: Presenting plural perspectives* (pp. 127–148). Springer. http://dx.doi.org/10.1007/978-94-007-7155-0_8.

Meder, B., & Mayrhofer, R. (2017). Diagnostic reasoning. In M. R. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 433–458). New York: Oxford University Press.

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. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review, 121*, 277–301. http://dx.doi.org/10.1037/a0035944.

Mosteller, F., & Youtz, C. (1990). Quantifying probabilistic expressions. *Statistical Science, 5*, 2–34. http://dx.doi.org/10.1214/ss/1177012251.

Olson, M. J., & Budescu, D. V. (1997). Patterns of preference for numerical and verbal probabilities. *Journal of Behavioral Decision Making, 10*, 117–131. http://dx.doi.org/10.1002/(SICI)1099-0771(199706)10:2<117::AID-BDM251>3.0.CO;2-7.

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

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111–163. http://dx.doi.org/10.2307/271063.

Rapoport, A., Wallsten, T. S., Erev, I., & Cohen, B. L. (1990). Revision of opinion with verbally and numerically expressed uncertainties. *Acta Psychologica, 74*, 61–79. http://dx.doi.org/10.1016/0001-6918(90)90035-E.

Reagan, R. T., Mosteller, F., & Youtz, C. (1989). Quantitative meanings of verbal probability expressions. *Journal of Applied Psychology, 74*, 433–442. http://dx.doi.org/10.1037/0021-9010.74.3.433.

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. http://dx.doi.org/10.1027/1618-3169/a000298.

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive Psychology, 72*, 54–107. http://dx.doi.org/10.1016/j.cogpsych.2014.02.002.

Rehder, B., & Waldmann, M. R. (2017). Failures of explaining away and screening off in described versus experienced causal learning scenarios. *Memory & Cognition, 45*, 245–260. http://dx.doi.org/10.3758/s13421-016-0662-3.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology, 50*, 264–314. http://dx.doi.org/10.1016/j.cogpsych.2004.09.002.

Rottman, B. M. (2016). Searching for the best cause: Roles of mechanism beliefs, autocorrelation, and exploitation. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*. http://dx.doi.org/10.1037/xlm0000244.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review, 117*, 1144–1167. http://dx.doi.org/10.1037/a0020511.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464. http://dx.doi.org/10.2307/2958889?.

Simpson, R. H. (1944). The specific meanings of certain terms indicating differing degrees of frequency. *Quarterly Journal of Speech, 30*, 328–330. http://dx.doi.org/10.1080/00335634409381009.

Simpson, R. H. (1963). Stability in meanings for quantitative terms: A comparison over 20 years. *Quarterly Journal of Speech, 49*, 146–151. http://dx.doi.org/10.1080/00335636309382600.

Sirota, M., & Juanchich, M. (2012). To what extent do politeness expectations shape risk perception? Even numerical probabilities are under their spell! *Acta Psychologica, 141*, 391–399. http://dx.doi.org/10.1016/j.actpsy.2012.09.004.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science, 27*, 453–489. http://dx.doi.org/10.1016/S0364-0213(03)00010-7.

Stone, D. R., & Johnson, R. T. (1959). A study of words indicating frequency. *Journal of Educational Psychology, 50*, 224–227. http://dx.doi.org/10.1037/h0044812.

Teigen, K. H., & Brun, W. (2003). Verbal expressions of uncertainty and probability. In D. Hardman (Ed.), *Thinking: Psychological perspectives on reasoning, judgment and decision making* (pp. 125–145). New York, NY: Wiley.

Thompson, W. C., & Newman, E. J. (2015). Lay understanding of forensic statistics: Evaluation of random match probabilities, likelihood ratios, and verbal equivalents. *Law and Human Behavior, 4*, 332–349. http://dx.doi.org/10.1037/lhb0000134.

Trueblood, J. S., & Busemeyer, J. R. (2011). A quantum probability account of order effects in inference. *Cognitive Science, 35*, 1518–1552. http://dx.doi.org/10.1111/j.1551-6709.2011.01197.x.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131. http://dx.doi.org/10.1126/science.185.4157.1124.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11*, 192–196. http://dx.doi.org/10.3758/BF03206482.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General, 121*, 222–236. http://dx.doi.org/10.1037/0096-3445.121.2.222.

Wallsten, T. S., & Budescu, D. V. (1995). A review of human linguistic probability processing: General principles and empirical evidence. *The Knowledge Engineering Review, 10*, 43–62. http://dx.doi.org/10.1017/S0269888900007256.

Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General, 115*, 348–365. http://dx.doi.org/10.1037/0096-3445.115.4.348.

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science, 39*, 176–190. http://dx.doi.org/10.1287/mnsc.39.2.176.

Wallsten, T. S., Fillenbaum, S., & Cox, J. A. (1986). Base rate effects on the interpretations of probability and frequency expressions. *Journal of Memory and Language, 25*, 571–587. http://dx.doi.org/10.1016/0749-596X(86)90012-4.

Weber, E. U., & Hilton, D. J. (1990). Contextual effects in the interpretations of probability words: Perceived base rate and severity of events. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 781–789. http://dx.doi.org/10.1037/0096-1523.16.4.781.

Wixted, J. T., & Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory & Cognition, 25*, 731–739. http://dx.doi.org/10.3758/BF03211316.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*, 338–353. http://dx.doi.org/10.1016/S0019-9958(65)90241-X.

Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning—I. *Information Sciences, 8*, 199–249. http://dx.doi.org/10.1016/0020-0255(75)90036-5.

Ziegler, A., Hadlak, A., Mehlbeer, S., & Konig, I. R. (2013). Comprehension of the description of side effects in drug information leaflets—A survey of doctors, pharmacists and lawyers. *Deutsches Ärzteblatt, 110*, 669–673. http://dx.doi.org/10.3238/arztebl.2013.0669.

Zimmer, A. C. (1983). Verbal vs. numerical processing of subjective probabilities. *Advances in Psychology, 16*, 159–182. http://dx.doi.org/10.1016/S0166-4115(08)62198-6.