Sequential Diagnostic Reasoning with Verbal Information

Björn Meder (meder@mpib-berlin.mpg.de)
Max Planck Institute for Human Development, Center for Adaptive Behavior and Cognition (ABC)
Lentzeallee 94, 10495 Berlin, Germany

Ralf Mayrhofer (rmayrho@uni-goettingen.de)
Department of Psychology, University of Göttingen
Gosslerstrasse 14, 37073 Göttingen, Germany

Abstract

In sequential diagnostic reasoning, the goal is to infer the probability of a cause event from sequentially observed effects. Typically, studies investigating such tasks provide subjects with precise quantitative information regarding the strength of the relations between causes and effects. By contrast, we examined people's performance when this information is communicated through qualitative, rather vague verbal terms (e.g., "*X* occasionally causes symptom *A*"). We conducted an experiment in which we compared subjects' judgments with a Bayesian model whose predictions were derived using numeric equivalents of various verbal terms from an unrelated study with different subjects. We found a remarkably close correspondence between subjects' diagnostic judgments based on verbal information and the model's predictions, as well as compared to a matched control condition in which information was presented numerically. Additionally, we observed interindividual differences regarding the temporal weighting of evidence.

Keywords: Sequential diagnostic reasoning; verbal reasoning; causal inference; Bayesian models; recency effects; linguistic probability terms; evidence accumulation

Introduction

In diagnostic reasoning, the goal is to infer the probability of a cause event from observing its effects. The characteristic feature of *sequential* diagnostic reasoning is that multiple pieces of evidence are observed at different points in time. For instance, a doctor whose aim is to infer the cause of a patient's symptoms may take a blood sample and order different diagnostic tests. The test results may come in distributed over time, with each result potentially providing evidence for different diseases. Thus, sequential diagnostic reasoning requires keeping track of the evidence and the hypotheses under consideration.

We investigated different aspects of sequential diagnostic reasoning. Theoretically, we considered different ways in which such tasks can be modeled. For instance, standard probability calculus (e.g., Bayes's rule) is not sensitive to the temporal dynamics of evidence accumulation. Yet, there are ways to incorporate temporal weighting of evidence into probabilistic models of diagnostic reasoning and to model its potential influence on people's inferences.

Empirically, we were interested in investigating diagnostic reasoning with verbal information. In many real-world situations, everyday language is used to communicate probability or frequency information (Budescu & Wallsten, 1995; Teigen & Brun, 2003). For example, we might find it unusual if a doctor told us that the probability of a particular disease causing some symptom is 66%. By contrast, a statement such as "disease *X* frequently causes symptom *A*" may feel more natural, despite the apparent lack of preciseness (Wallsten, Budescu, Zwick, & Kemp, 1993).

Although using verbal probability terms is common in many real-world situations, they do not easily fit with computational accounts of cognition. As a consequence, in most behavioral studies subjects are provided with precise quantitative information (e.g., Meder, Mayrhofer, & Waldmann, 2009). By contrast, we investigated reasoning with verbal information by using the numeric equivalents of linguistic terms (Bocklisch, Bocklisch, & Krems, 2012) to derive model predictions. This allowed us to examine people's capacity to make diagnostic inferences in the absence of quantitative information and to compare their judgments to different accounts, including variants of Bayesian and linear models. To test for the temporal weighting of information, we varied the testing conditions by manipulating whether all evidence obtained so far was directly available when making a judgment or had to be partially retrieved from memory.

Modeling Sequential Diagnostic Reasoning

The characteristic feature of sequential diagnostic reasoning is that different pieces of evidence are acquired step by step. Consider the causal model shown in Figure 1a. There are two (mutually exclusive) cause events, *X* and *Y*; each can generate effects *A*, *B*, *C*, and *D*. In our experiment the cause variables were two chemical substances and the effects were different symptoms caused by these substances. The symptoms were observed sequentially and the goal was to infer whether *X* or *Y* caused them. How can such inferences be formally modeled?

Standard Model: "Simple" Bayes

Let S denote a set of symptoms $\{S_1, \ldots, S_T\}$, 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:

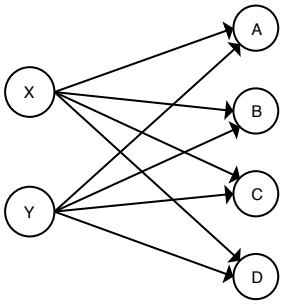
$$P(Y|S) = \frac{P(S|Y)P(Y)}{P(S|Y)P(Y) + P(S|X)P(X)} \qquad (1)$$

where $P(S|Y)$ denotes the likelihood of the symptoms given cause Y, $P(Y)$ is the base rate of cause Y, and $P(S|X)$ and $P(X)$ denote the corresponding estimates for the alternative cause.

We considered only situations in which X and Y were equally likely a priori, thus $P(X) = P(Y) = .5$. In this case, Eq. (1) simplifies to

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

Thus, the posterior probability of Y is a function of the likelihood of the set of symptoms S under each of the two hypotheses X and Y.



a) Causal model

b) Relations and presentation formats

Likelihoods	Numeric	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"

Figure 1: (a) Causal structure used in our diagnosis task, and (b) strength of the individual links (likelihoods) in the numeric and verbal formats used in the experiment.

Temporal Weighting of Evidence: "Memory" Bayes

For the simple Bayes model, the temporal order in which observations are made does not matter: The resulting probabilities are the same regardless of whether beliefs are updated sequentially according to the individual symptoms or conditional on all symptoms at once.

However, we were also interested in modeling the sequential dynamics of evidence accumulation. For instance, diagnostic inferences can be influenced by memory limitations, such as the (partial) neglect of earlier obtained evidence. To model the influence of time, we applied the log odds form of Bayes's rule to the target inference:

$$\varphi = \log\frac{P(Y|S)}{P(X|S)} = \log\frac{P(S|Y)}{P(S|X)} + \log\frac{P(Y)}{P(X)} \tag{3}$$

Assuming both hypotheses are equally likely a priori, we can omit the prior odds from the derivation and expand the likelihood odds by summing over the sequence of symptoms $S_1, \ldots, S_T$ given their conditional independence:

$$\varphi = \sum_{t=1}^{T} \log\frac{P(S_t|Y)}{P(S_t|X)} \tag{4}$$

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:

$$P(Y|S) = \frac{1}{1 + e^{-\varphi}} \tag{5}$$

This equation is mathematically equivalent to the standard form of Bayes's rule for the posterior probability of Y given the set of symptoms S as shown in Eq. (2).

Importantly, the log-odds form allows us to introduce an exponential decay parameter δ that controls the weighting of symptoms in the course of their presentation (Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). Therefore, we replace Eq. (4) by

$$\varphi = \sum_{t=1}^{T} \left[\log\frac{P(S_t|Y)}{P(S_t|X)}\right] e^{-\frac{T-t}{\delta}} \tag{6}$$

In the limit, if $\delta = \infty$, there is no decay and Eq. (6) reduces to Eq. (4). In this case, all symptoms are equally weighted and symptom order does not matter (as predicted by the simple Bayes model). The closer δ is to 0, the more weight is given to the current symptom. If $\delta = 0$, the posterior probability depends on only the most recent symptom, yielding an inference strategy that is completely ignorant of past information (e.g., an agent without memory). Thus, "memory" Bayes can be used to model recency effects (Hogarth & Einhorn, 1992; Trueblood & Busemeyer, 2011).

Mapping Verbal Terms to Numbers

A key question of our research was how accurately people reason with verbal information, absolutely and relative to situations in which quantitative information is available. Answering this requires translating verbal expressions into a numeric representation that can be used to derive precise model predictions for the verbal reasoning task.

We used numeric equivalents of verbal expressions from a study by Bocklisch et al. (2012). They asked subjects to provide numeric estimates for different verbal expressions in a frequency format (e.g., It is *frequently* necessary to work at a rapid pace means "in X of 100 work tasks/cases"). This mapping of words to numbers provided the basis for our empirical study, in which we used the four verbal expressions "infrequently", "occasionally", "frequently", and "almost always" to convey the strength of the cause–effect relations.[1] The corresponding numeric mean estimates were 19%, 29%, 66%, and 88% (Figure 1b).

These estimates were used to derive posterior probabilities of the causes given the symptoms via Bayes's rule [Eq. (1)], which served as normative benchmarks for evaluating subjects' diagnostic judgments. Note that the numeric equivalents were elicited from a different, unrelated sample from the one used in our study.

[1] 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.

Experiment

The main goal of our study was to investigate sequential diagnostic reasoning with verbal information and compare different presentation formats with respect to the temporal weighting of evidence.

The first factor we manipulated 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 relations was conveyed through four verbal terms ("infrequently", "occasionally", "frequently", "almost always"). In the *numeric* 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 numeric condition subjects learned that the probability of *X* causing *A* is 88% (see Figure 1b).

With the second manipulation we aimed to investigate possible influences of temporal weighting on diagnostic judgments (i.e., recency effects). In the *single-symptom* condition, only the current symptom was visible on the computer screen when participants made the diagnostic judgment. In the *all-symptoms* condition, the full set of symptoms reported *so far* was visible on the screen when they made a diagnosis.[2] The rationale behind this manipulation was that there might be a tendency to overweight the currently presented symptom when previously obtained evidence needs to be recalled from memory.

Method

Participants One hundred fifty-six students (103 women; M_{age}=23.4 years) 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 per hour.

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. The instructions asked subjects to take the role of a doctor being responsible for the workers at two chemical plants. At one plant workers may come in contact with the substance "Altexon"; at the other they may have contact with "Zyroxan". Each of these substances can cause four symptoms: dizziness, fever, headache, and vomiting. The assignment of labels to causes (substances) and effects (symptoms) was randomized.

Subjects were informed that their task would be to diagnose a series of workers who had had contact with either of the two substances. The instructions explicitly stated that accidents were equally likely to happen in each of the plants (i.e., the base rate of each cause was 50%). Subjects were also told that the patients would report their symptoms sequentially.

The experiment consisted of two phases: a learning phase, in which subjects learned the strengths of the individual causal relations, and a test phase, in which subjects were sequentially presented with symptoms of different patients and had to make a diagnostic judgment after each symptom.

Figure 1b illustrates the strengths of the relations 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 often the substance causes the symptom. In the verbal condition, possible answers were "infrequently", "occasionally", "frequently", and "almost always". In the numeric condition, the corresponding answers were 19%, 33%, 66%, and 88%. After making an estimate, subjects received feedback regarding the actual relation. The eight relations were presented block-wise, with the order randomized within a block. To proceed to the test phase, subjects needed to answer two consecutive blocks correctly (or pass 12 blocks in total).

In the test 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. All judgments were given on an 11-point scale from 0 to 100, with the endpoints labeled as "The patient definitely had contact with Altexon" and "The patient definitely had contact with Zyroxan".

Table 1 shows the six symptom sequences together with the posterior probabilities derived using the likelihoods shown in Figure 1b, assuming $P(X) = P(Y) = .5$. Additionally, we presented the six symptom sequences that entailed identical posterior probabilities for *X* (e.g., $P(Y|A–D–C) = P(X|D–A–B)$ such that diagnoses were counterbalanced. Thus, each subject saw 12 sequences in total. The corresponding pairs were later recoded and aggregated.

The test trials were administered in random order. After the test phase, we tested subjects again with respect to the strength of the individual substance–symptom relations (as learned in the learning phase) by presenting an additional

Table 1: Test trials with sequentially presented symptoms.

Posterior Probability	Symptom sequence					
	A–D–C	*D–A–C*	*B–C–A*	*C–B–A*	*A–C–D*	*C–A–D*
$P(Y\|S_1)$	.18	.82	.31	.69	.18	.69
$P(Y\|S_1, S_2)$	.50	.50	.50	.50	.33	.33
$P(Y\|S_1, S_2, S_3)$	.69	.69	.18	.18	.69	.69

Note. Numbers refer to the posterior probability of cause *Y* given a set of symptoms *S* according to the simple Bayes model.

[2] Note, *all* symptoms does not mean that subjects were presented with all symptoms of a trial at each time, but with all symptoms that were relevant to the current judgment. 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 $\{S_1\}, \{S_2\}, \{S_3\}$.

block of the learning phase (without feedback). This served as a manipulation check to ensure that subjects still remembered the relations between substances and symptoms.

Design Subjects were randomly assigned to one of the 2 (numeric vs. verbal) × 2 (single vs. all symptoms) between-subjects conditions. Within each subject, we aggregated over the (recoded) judgments within the counterbalanced pairs of trials, yielding 6 (trials) × 3 (symptoms) within-subject conditions with judged $P(Y|S)$ as dependent measure.

Results and Discussion

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 is 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 $N = 112$).

Overall Fit Figure 2 shows subjects' mean diagnostic judgments for the different symptom sequences along with the posterior probabilities derived from the simple Bayes model. A first inspection of the data indicates that subjects' judgments were remarkably accurate, with estimates being close to the true posteriors. This was the case regardless of whether information was provided in a verbal or numeric format (especially in the all-symptoms condition; see left-hand side of Figure 2). Thus, subjects were capable of making pretty accurate inferences when reasoning with verbal information. This close correspondence is particularly remarkable because the numeric equivalents of the verbal terms were taken from a different sample of subjects who participated in an unrelated study (Bocklisch et al., 2012).

Before conducting the model-based analysis, we ran a mixed analysis of variance with the 2 (numeric vs. verbal) × 2 (single vs. all symptoms) conditions as between-subjects factors and the 6 (trials) × 3 (symptoms) conditions as within-subject factors. The key result of this analysis was that there was no main effect of presentation format, $F(1, 108) < 1$, a weak effect of the single- vs. all-symptom presentation manipulation, $F(1, 108) = 3.1$, $p = .08$, $\eta_p = .03$, and no interaction ($F < 1$).

To evaluate subjects' overall accuracy we computed the correlation and mean squared error (*MSE*) between the empirical judgments and the posterior probabilities derived from the simple Bayes model (note that no fitting is involved here). To address if symptoms are weighted differently in sequential reasoning, we fitted the decay parameter δ of the "memory" Bayes model to the data (separately for each condition, using the *MSE* as fitting criterion).[3] The relative size of the decay parameter δ in the single- vs. all-symptoms condition gives an idea of whether the testing

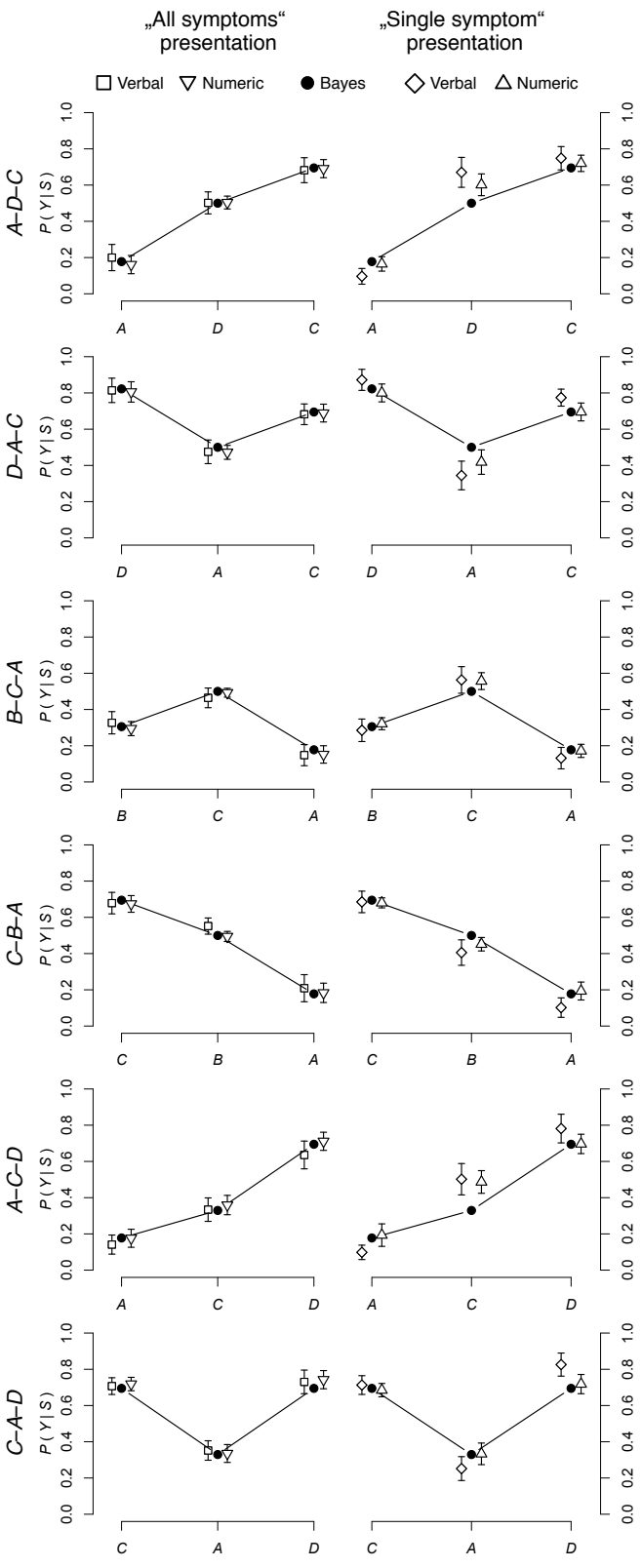


Figure 2: Mean diagnostic judgments (±95% CI) and predictions of the simple Bayes model. Rows represent the different trials (see Table 1), separately by testing procedure (all-symptoms vs. single-symptom presentation).

[3] For this purpose, we used a grid search over a plausible set of values for δ between 1e–10 and 1e+10.

procedure influences people's judgments, in particular whether there is a tendency to neglect previous evidence when only the current symptom is shown when performing a diagnostic judgment. This should result in lower values of the decay parameter δ for the single-symptom condition relative to the all-symptoms condition.[4]

Table 2: Fits of the "simple" and "memory" Bayes models.

Format	Symptoms	Simple Bayes		Memory Bayes		
		r	*MSE*	*r*	*MSE*	δ
Verbal	All	.991	.0008	.991	.0009	>1e+10
	Single	.952	.0089	.983	.0036	2
Numeric	All	.996	.0004	.996	.0004	40
	Single	.971	.0028	.988	.0014	4.5

Table 2 shows the fits of the two models. Overall, both the (high) correlations and the (low) *MSE* indicate that the models' predictions fit well with subjects' judgments. In the all-symptoms conditions, the fit for simple Bayes was almost perfect (r = .991 and r = .996, respectively), mirroring that for 35 of 36 (6 trials × 3 symptoms × 2 formats [verbal vs. numeric]) data points, the model's predictions fell inside the 95% confidence interval.

The results also indicate 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 (cf. Figure 2). Here, in both the verbal and the numeric condition, lower values for δ were obtained than in the all-symptoms conditions (see Table 2). Consistent with this finding, for these conditions the memory Bayes model achieved a higher fit than the simple Bayes model, in terms of both the correlation and the *MSE*. This result indicates that subjects were more likely to overweight the current evidence when previous symptoms had to be recalled from memory.

Model-Based Clustering Temporal weighting of cumulative evidence might not be due to the characteristics of the task or the reasoning context alone but might also result from interindividual differences or strategies. We therefore explored if it is possible to identify homogenous subgroups of subjects differing with respect to their temporal weighting of symptoms (i.e., that differ in the δ parameter).

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 homogenous with respect to the model predictions, and second, estimating the best-fitting δ parameter for each cluster. This problem can be approximately solved by a recursive algorithm that starts with a random assignment of subjects to clusters and then iterates over two steps, namely, fitting and re-assignment, until no subject changes cluster.[5]

[4] Remember that in the limit, if δ = ∞ there is no decay; if δ = 0 the posterior probability depends on only the most recent symptom.

[5] More specifically, the algorithm proceeds as follows: (i) Given the current assignments of participants to clusters, find the δ parameter for each cluster that minimizes the *MSE* of the model predictions with respect to the average response profile of the subjects within the cluster. (For this purpose, we used a grid search over a plausible set of value for δ.) (ii) Given the model predictions for the different clusters, reassign subjects to a cluster such that the correlation between the individual response profile and the model prediction is maximized. Then, iterate through (i) and (ii) until no participant changes cluster anymore.

Table 3: Results of the model-based clustering.

Format	Symptoms	Cluster 1				Cluster 2			
		δ	*n*	*r*	*MSE*	δ	*n*	*r*	*MSE*
Verbal	All	∞	28	.991	.0008	–	–	–	–
	Single	∞	15	.994	.0014	0	12	.993	.0047
Numeric	All	∞	24	.997	.0003	1.5	3	.969	.0056
	Single	85	21	.993	.0010	0.6	9	.989	.0017

Note. δ = ∞ means that the estimate is greater than 1e+10; in this case there is essentially no difference from the predictions of the simple Bayes model. δ = 0 means that the estimate is smaller than 1e–10; in this case there is essentially no difference from the prediction of Bayes's rule taking into account *only* the currently presented symptom.

We applied this procedure to each of the four conditions; the results are shown in Table 3. The verbal all-symptoms condition yielded only one cluster as a solution, whereas the other three conditions yielded stable two-cluster solutions. Remarkably, in each condition the majority of subjects were assigned to a cluster that is best represented by a very high δ parameter of the memory Bayes model. Essentially, this means that these subjects are best described by a prediction profile that is almost identical to the predictions of the simple Bayes model. In the single-symptom conditions, however, a substantial proportion of people were best described by a quite low δ parameter, meaning that their diagnostic judgments were almost exclusively determined by the currently presented symptom. Taken together, the clustering results strengthen the findings we already obtained by the overall fitting of the data.

Linear Models of Diagnostic Judgment In our study, people's diagnostic judgments corresponded strongly to the predictions of Bayes's rule. Can alternative models approximate these predictions? We here consider one alternative class of models, namely, weighted-additive (WADD) approaches. From this view, the cause event is inferred using an average (i.e., linear) combination of symptom weights:

$$P(Y|S) = \frac{1}{T}\sum_{t=1}^{T} w_{S_t} \qquad (7)$$

where t is the current symptom and T is the total number of symptoms observed so far.

We tested three different linear models that make different assumptions regarding the decision weights. The simplest model, *tallying*, simply counts symptoms. In our scenario, symptoms *A* and *B* are more likely to be generated by *X*, whereas *C* and *D* are more likely to be generated by *Y*. Given a set of symptoms, one simply tallies the evidence. For instance, given the sequence *A*–*C*–*D*, two of the three symptoms provide evidence for *Y*; accordingly, the resulting

estimate would be 2/3. Note that this result is very close to the true probability, which is .69 in this case.

The second linear model assumes that the decision weights reflect the strength of the cause–effect relations; we therefore call it *likelihood WADD*. This model simply sums over the likelihoods and normalizes the result by dividing it by the number of presented symptoms. Given the sequence *A–C–D*, this model would predict that the probability of *Y* is .58 [(.19 +.66 +.88)/3], which for this sequence is quite close to the true probability of .69.

Finally, we examined the predictions of an "optimal" WADD model by fitting the weights to the data, using *MSE* minimization as a criterion. This model essentially serves as a benchmark, as it provides the best fit given the functional form of the model (linear combination) and the data.

Table 4: Fits of the linear models.

Format	Symptoms	Tallying		Likelihood WADD		Optimal WADD	
		r	*MSE*	*r*	*MSE*	*r*	*MSE*
Verbal	All	.861	.0278	.901	.0093	.901	.0085
	Single	.817	.0320	.856	.0242	.857	.0197
Numeric	All	.864	.0266	.896	.0106	.898	.0094
	Single	.848	.0285	.880	.0108	.881	.0102

The results (Table 4) show that all linear models achieved a respectable fit, but none could match the Bayesian models. These results speak against the idea that our subjects used a linear-additive strategy to make judgments.

General Discussion

Although verbal terms such as "infrequently", "occasionally", and "frequently" 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 numeric information in order to compare their behavior and inferences to the predictions of computational models, which typically also require numeric 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 reliable quantitative information. Using a sequential diagnostic reasoning task, we observed that people's inferences were surprisingly accurate when information on cause–effect relations was conveyed merely through linguistic terms. In fact, performance was almost indiscernible from a control condition in which subjects were provided with numeric information. The fact that we took the numeric equivalents from a different study (Bocklisch et al., 2012) supports research showing that the interpretation of linguistic frequency terms is relatively stable across populations (Mosteller & Youtz, 1990).

Generally, subjects' diagnostic judgments closely resembled the prediction of a simple Bayes model that operates on matched numeric values. 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 numeric point estimates (e.g., mean of an elicited frequency term distributions) but can also operate on full distributions (e.g., fitted Beta distributions).

Furthermore, we investigated the temporal weighting of evidence. We found that symptoms were equally weighted when all relevant symptoms were available during judgment, but we also observed a neglect of previous evidence when only the current symptom was present. Model-based cluster analyses revealed that this was due to a subgroup of subjects who considered only the current symptom, whereas most people took into account all evidence in a normative fashion. Overall, our results contrast with views that consider human probabilistic reasoning as flawed and error prone.

Acknowledgments

Both authors contributed equally to this research. We thank Franziska Bocklisch, Georg Jahn, Felix Rebitschek, and Michael Waldmann for helpful comments on this project. This research was supported by grants Wa 621/22 and Me 3717/2 from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "New Frameworks of Rationality" (SPP 1516).

References

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.

Budescu, D. V., & Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. *Psychology of Learning and Motivation*, *32*, 275–318.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*, 1–55.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2009). A rational model of elemental diagnostic inference. In *Proceedings of the 31st annual conference of the Cognitive Science Society* (pp. 2176–2181). Austin, TX: Cognitive Science Society.

Mosteller, F., & Youtz, C. (1990). Quantifying probabilistic expressions. *Statistical Science*, *5*, 2–34.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*, 453–489.

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*. New York, NY: Wiley.

Trueblood, J., & Busemeyer, J. R. (2011). A quantum probability account of order effects in inference. *Cognitive Science*, *35*, 1518–1552.

Wallsten, T. S., Budescu, D. V., Zwick, R., & Kemp, S. M. (1993). Preferences and reasons for communicating probabilistic information in numerical or verbal terms. *Bulletin of the Psychonomic Society*, *31*, 135–138.