ELSEVIER

Cognition 83 (2002) 81–112

COGNITION

www.elsevier.com/locate/cognit

Probability judgment in hierarchical learning: a conflict between predictiveness and coherence

David A. Lagnado*, David R. Shanks

Department of Psychology, University College London, Gower Street, London WC1E 6BT, UK

Received 6 March 2001; received in revised form 15 June 2001; accepted 14 November 2001

Abstract

Why are people's judgments incoherent under probability formats? Research in an associative learning paradigm suggests that after structured learning participants give judgments based on predictiveness rather than normative probability. This is because people's learning mechanisms attune to statistical contingencies in the environment, and they use these learned associations as a basis for subsequent probability judgments. We introduced a hierarchical structure into a simulated medical diagnosis task, setting up a conflict between predictiveness and coherence. Thus, a target symptom was more predictive of a subordinate disease than of its superordinate category, even though the latter included the former. Under a probability format participants tended to violate coherence and make ratings in line with predictiveness; under a frequency format they were more normative. These results are difficult to explain within a unitary model of inference, whether associative or frequency-based. In the light of this, and other findings in the judgment and learning literature, a dual-component model is proposed. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Probability judgment; Hierarchical structure; Associative learning; Frequency format; Conjunction fallacy

1. Introduction

Uncertainty is a pervasive feature of the world around us. Whether it is a doctor learning to diagnose a new disease, an investor tracking the prices of shares, or a tipster picking winners at the races, people are continually acquiring imperfect information about their environment, and making judgments on this basis. Intuitively one would expect there to be considerable interplay between the *learning* of probabilistic information, and the *judgment* made under this uncertainty, but these aspects of cognition tend to be studied in separate research paradigms. Thus, questions about how we learn about the structure of the world

* Corresponding author.
E-mail address: d.lagnado@ucl.ac.uk (D.A. Lagnado).

PII: S0010-0277(01)00168-8

are framed in terms of our sensitivity to contingency relations between events (Cheng & Novick, 1992; Shanks, 1995), whereas questions about our probability judgments are analyzed with reference to a normative probability model (e.g. Tversky & Kahneman, 1983). In line with recent attempts to unify these concerns (López, Cobos, Caño, & Shanks, 1998), this paper will examine how the mechanisms that underlie learning affect subsequent judgments of probability. In particular we will focus on the *coherence* of people's judgments – whether they conform to the laws of probability – after exposure to a dynamic learning environment.

1.1. Why are probability judgments incoherent?

Given the intensity of recent debates about *whether* people's judgments violate the laws of probability (e.g. see Gigerenzer, 1996; Kahneman & Tversky, 1996; Koehler, 1996; Vranas, 2000), it is not surprising that the question of *why* they might do so has taken a backseat. However, despite a general consensus that the charge of widespread 'irrationality' is premature, the incoherence of our probability judgments remains a robust empirical finding. Indeed, the numerous experiments that purport to show that judgmental biases are significantly reduced by presenting problems in a frequency format (e.g. Cosmides & Tooby, 1996; Fiedler, 1988; Gigerenzer & Hoffrage, 1995) serve also to highlight the reality of these biases in a probability format. Even if one disputes the appropriateness of the laws of probability as a normative model in such contexts (Hertwig & Gigerenzer, 1999), the question remains as to what people are doing when they exhibit judgmental biases under the probability format.

1.2. An associative model of learning and judgment

There is a growing body of research in the area of human associative learning that suggests an explanatory framework within which to answer such questions (e.g. Gluck & Bower, 1988; López et al., 1998; Price & Yates, 1995; Shanks, 1990). Associative theories of learning are concerned primarily with how people detect and assess contingency relationships between events in their environment. Based on earlier research in animal conditioning (e.g. Dickinson, 1980), associative theories maintain that learning consists of the formation of psychological associations between cue–outcome pairs, and that the strength of these associations is determined by the extent to which the cue is a valid predictor of the outcome. On most associative accounts learning is mediated through the operation of a discrepancy-based mechanism such as the Rescorla–Wagner (RW) rule (Rescorla & Wagner, 1972), which is formally equivalent to the delta rule used in many connectionist networks (Stone, 1986). This rule adjusts the associative weight between cue–outcome pairs on a trial-by-trial basis, by minimizing predictive error. That is, associative bonds are strengthened when the presence of the cue successfully predicts the occurrence of the outcome and weakened when the outcome it predicts fails to occur.

As well as providing a model for *experiential* learning, the associative framework provides an account of the judgments reached as a result of this learning. The basic idea is that the associative strengths formed in learning serve as the basis for judgments about the contingency, or predictive strength, between events. Thus, people will rate the contingency or causal relatedness of a cue–outcome pair according to how well the cue predicts the outcome. Despite the simplicity of the associative account, it can explain a wide variety of empirical findings in human learning (see Shanks, 1995). Further, it has been demonstrated (Cheng, 1997) that the associative strengths generated by the RW rule will converge on the probabilistic contrast ΔP, an objective measure for the statistical contingency between pairs of events.

Due to a predominant interest in how we learn about causality, most human learning research uses contingency judgments as the dependent response variable. However, a fundamental assumption of the associative approach is that these learned associations also underwrite a range of other kinds of response and judgment. In particular, a number of studies have looked at the influence of learned associations on subsequent probability judgment (e.g. Gluck & Bower, 1988; Price & Yates, 1995; Shanks, 1990). In a typical experiment participants first undergo a training phase in which they learn to associate cues (e.g. symptoms) with outcomes (e.g. diseases). After this training, they are required to estimate the probability of outcome given cue for various cue–outcome pairs. A major finding in such research is that people tend to make probability judgments that are based on the predictive relation between cue and outcome, rather than a normative representation of probability. That is, when asked to estimate the probability of a disease given a symptom, participants respond with an evaluation of the degree to which the symptom predicts the disease, rather than the true conditional probability of the disease given the symptom (the two notions are readily confused, but, as we argue below, can lead to quite different measures). The associative model has a ready explanation for this finding – in the training phase of these experiments participants form appropriate associations between events, and under test they use these as a basis for their probability judgments.

1.3. The distinction between predictiveness and conditional probability

Before outlining some of the evidence that supports this associative claim, it is worth noting, at a more general level, the relation between the notions of predictiveness and conditional probability. As mentioned above, an objective measure of the degree to which one event predicts another is given by the *statistical contingency* ΔP (see Allan & Jenkins, 1980). With a single cue C and outcome O this corresponds to the *difference* between the probability of the outcome given the cue, $P(O|C)$, and the probability of the outcome in the absence of the cue, $P(O|\sim C)$:

$$\Delta P = P(O|C) - P(O|\sim C) \tag{1}$$

In other words, the extent to which a cue is a good predictor of an outcome is measured by the extent to which the presence of the cue *raises* the probability of the outcome. Thus, the notion of predictiveness is a function of two conditional probabilities, but not itself a conditional probability.

In spite of this conceptual difference, in many situations the degree to which a cue predicts an outcome will correlate with the probability of the outcome given the cue. At the extreme, if the outcome can only occur in the presence of a particular cue, then the two objective measures will correspond exactly, i.e. if $P(O|\sim C) = 0$, then $\Delta P = P(O|C)$. More commonly, with several potential cues, there will be a positive correlation between these measures. That is, the higher the conditional probability $P(O|C)$, the higher the contingency ΔP between O and C, and vice-versa.

There are, however, certain contexts in which these two measures will differ quite substantially. One clear case is when a cue has no predictive value for an outcome (the two events are independent), but the outcome has high probability anyway. For example, the probability that 'Red Rum wins the race' (R) given that 'the jockey wears blue' (B) may be identical to the probability that 'Red Rum wins the race' given that 'the jockey does not wear blue' ($\sim B$), and both may be high, whereas the contingency between B and R is equal to zero. A more subtle case occurs when a cue differs in its predictive strength for each of two outcomes, but their respective conditional probabilities are equivalent. Such a structure is exploited in the base rate experiments to be discussed in the next section.

To summarize, in many contexts the predictiveness (of outcome from cue) and the conditional probability (of outcome given cue) may yield similar measures, but in certain circumstances these will differ. Of empirical interest here is how sensitive our inference mechanisms are to this distinction. That is, given that the notions are logically distinct, are they kept psychologically distinct? As mentioned earlier, associative theorists argue that people are prone to conflate the two, and use the same primitive – a learned associative strength – for both kinds of judgment. The classic demonstration of this in the human learning literature is the phenomenon of base rate neglect.

1.4. Base rate neglect

In a learning analogue of Tversky and Kahneman's studies of base rate usage (Tversky & Kahneman, 1982), people are required to rate the probability of each of two diseases given a particular target symptom (Gluck & Bower, 1988; Medin & Edelson, 1988; Shanks, 1990). The learning environment is arranged so that these two conditional probabilities are equal, but the overall probability (the base rate) of one disease is high and the other low. Given this structure, the target symptom is a better predictor of the rare disease than of the common disease. The results obtained with this task have remained constant across numerous replications (e.g. Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Nosofsky, Kruschke, & McKinley, 1992; Shanks, 1991): people rate the conditional probability of the rare disease higher than the common disease. Thus, the judgments in these experiments reflect the degree to which the target symptom *predicts* the disease, and not simply the conditional probability of the disease given the symptom. This example of base rate neglect is readily accommodated within the associative learning paradigm, and is prima facie evidence in its favor.

1.5. Rule-based accounts

The main alternative to an associative model of human contingency judgment is given by *rule-based* accounts (e.g. Allan, 1993). On such a view, people are assumed to encode event frequencies during learning, and then use these representations to compute the degree of contingency between events. Here again a widely accepted normative standard is the statistical contingency metric ΔP. Most rule-based accounts assume that people arrive at their contingency judgments by explicitly computing this statistic (or some

	Outcome Present	Outcome Absent
Cue Present	*a*	*b*
Cue Absent	*c*	*d*

Fig. 1. Identification of cells in 2×2 contingency table.

variation of it) from the appropriate event frequencies. Thus, in the simple case with single cue and outcome, people are supposed to register frequencies for all four possible cue–outcome combinations (as shown in the 2×2 contingency table in Fig. 1). To reach a contingency judgment, they combine these frequencies via an explicit mental version of the ΔP rule (see Eq. (1) above). In terms of the notation in Fig. 1, $\Delta P = a/(a + b) - c/(c + d)$.

There are, therefore, two critical differences between the rule-based and the associative accounts. First, on the rule-based account people learn by updating event frequencies, whereas on the associative account they learn by updating associative connections. Second, on the rule-based account people make judgments via the *explicit computation* of the statistical contingency (or some variation of it) from the event frequencies. In contrast, on the associative account judgments are based on the learned associations. Whilst at asymptote these associative strengths will conform to ΔP, no explicit computation of this statistic is involved.

Although rule-based models can accommodate much of the data on contingency judgment (Allan, 1993), they have difficulty with certain findings (see Price & Yates, 1995; Shanks, López, Darby, & Dickinson, 1996). In particular, on such an account it is not clear why people should err in their conditional probability judgments but not their contingency judgments. If people are mentally comparing $P(O|C)$ and $P(O| \sim C)$ to reach a judgment of contingency, it is puzzling why they have difficulty in supplying $P(O|C)$ when requested. For the associative theory there is no puzzle, because probability judgments are based on learned associative strengths, and not on explicit representations of probability. This shortcoming of the rule-based account is highlighted in the learning studies on base rate neglect, in which people are often accurate in their contingency judgments whilst being biased in their probability judgments. It will also prove of relevance to the probability judgments to be investigated in this paper.

1.6. Violations of the laws of probability

Whilst base rate neglect is clearly a judgmental bias, it is arguable whether it actually constitutes a violation of the laws of probability. After all, these laws primarily concern the relationship between probability judgments, and place no constraints on the exact values assigned to any particular event or outcome (other than 1 to a logical truth). When the objective probability of event A is higher than event B, it is clearly *incorrect* to assign B a higher value than A, but it is not necessarily *incoherent* to do so. This will depend on how these assignments relate to one's other judgments. This distinction is best illustrated by considering a hypothetical betting situation.

Suppose that you are prepared to bet in accordance with your probability judgments about some future event (e.g. the result of an election, the winner of a horse race, etc.). If your judgments are in fact inaccurate, then someone can arrange bets with you so that you are *more likely* to lose money than to win it. They simply bet against the outcome that you (incorrectly) rate as probable, and/or bet on the outcome that you (incorrectly) rate as improbable. But they cannot *guarantee* that you lose. In contrast, if your judgments are incoherent, and violate the laws of probability, then someone can arrange bets with you so that you are *guaranteed* to lose money, irrespective of the result of the events bet upon.[1]

So, although the case of base rate neglect shows that in certain learning environments people give incorrect probability estimates, it does not demonstrate an explicit violation of the laws of probability. It is quite possible to have inaccurate probability judgments that are not thereby incoherent. Whilst keeping within the dynamic learning paradigm, we will examine whether people are prone to make sets of judgments that are explicitly incoherent.

1.7. The rule of extension

Fundamental to the laws of probability is the rule of extension, which states that if A is a subset of B, then $P(B) \geq P(A)$. This principle seems to accord well with common sense: if the occurrence of A entails the occurrence of B, but not vice-versa, then the chances of B occurring must be at least as high as the chances of A occurring. For example, the probability that a child has a contagious disease must be at least as high as the probability that they have German measles, because the set of contagious diseases includes German measles as a subset. The rule of extension has particular relevance when applied to knowledge systems with a hierarchical structure. Superordinate categories by definition include their subordinate categories, and thus a superordinate category can never be less probable than one of its subordinates. To illustrate, consider the subordinate category ‘British’ and the superordinate category ‘European’. For any individual (regardless of political orientation) the probability that they are British cannot be less than the probability that they are European.

In spite of the intuitive logic behind the rule of extension, Bar-Hillel and Neter (1993) found it to be violated in a set of one-shot, *described* problems. In a study typical of the heuristics and biases paradigm, participants received brief character profiles, and had to rank order a list of categories with respect to the probability that the character belonged to each category. This list included nested subordinate–superordinate pairs, and participants consistently ranked subordinate categories higher than their superordinates. Thus, on the basis of a particular student profile, they ranked ‘English Literature’ as a more likely subject category than ‘Humanities’, even though the latter included the former. Bar-Hillel and Neter termed these violations *disjunction errors*, and attributed them to the representativeness heuristic: people made their probability rankings according to the degree to which the category was representative of the character profile, rather than the degree to which the category was probable given the profile.

One criticism of this study (for another see Hertwig & Gigerenzer, 1999) concerns its

[1] This result is demonstrated by means of a *Dutch book argument* (e.g. see Finetti, 1980; Howson & Urbach, 1993; Osherson, 1995).

claim that the data support the representativeness heuristic over an explanation in terms of predictability. In addition to providing probability rankings, participants also ranked categories in terms of 'willingness to bet', 'inclination to predict', and 'suitability'. From the fact that there was no significant difference between any of these criteria, and in particular the systematic correlation between *suitability* and *probability* rankings, the authors concluded in favor of representativeness. But by parity of reasoning, the same argument applies to the observed correlation between *inclination to predict* and *probability* ranking. Indeed, out of the three ranking criteria, the *inclination to predict* correlated most highly with *probability*. So nothing in the data itself supports representativeness over an alternative account in terms of predictability. If anything, it suggests both may arise from the same underlying process. In sum, as well as demonstrating that people can violate the extension rule in one-shot judgment tasks, Bar-Hillel and Neter's study (unintentionally) lends some support to the idea that people interpret probability requests in terms of predictability.

2. A hierarchical learning task

The experiments reported in this paper introduce a simple hierarchical structure into the learning paradigm. This permits a natural test of the coherence of people's probability judgments, and of the associative conjecture that these judgments are based on learned associations rather than explicit representations of probability. Moreover, the hierarchical structure adds an extra dimension of realism into the learning task. Research in categorization has shown that category hierarchies are intrinsic to most knowledge domains, and play a critical role in classification and inference (Murphy & Lassaline, 1997; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). A better understanding of how people make probability judgments over hierarchically arranged content should complement such research.[2]

2.1. A conflict between coherence and predictiveness

In order to explore the hypothesis that probability judgments reflect estimates of predictiveness rather than conditional probability, we arranged the task environment so that the dictates of predictiveness conflict with those of coherence. There is, in fact, an important difference between these two kinds of standard. Whereas the standard of coherence, as furnished by the laws of probability, concerns relationships w*ithin* a set of judgments, the standard of predictiveness concerns the relationship *between* a judgment and the environment about which that judgment is made.[3] Thus, a judgment about the degree to which one event predicts another is appraisable with respect to the degree to which these events are

[2] The categorization literature includes some early analyses of the relation between cue validity and hierarchical category structure (e.g. Medin, 1983; Murphy, 1982). However, this work focused on the inadequacy of cue validity as a measure for distinguishing 'basic-level' categories, and did not explicitly consider how hierarchical structure can lead to an opposition between cue predictiveness and probabilistic coherence.

[3] This parallels the distinction made by Hammond (1996) between coherence and correspondence criteria.

actually related in the environment. This dependence on real world contingencies contrasts sharply with the content-independent nature of probabilistic coherence.

One consequence of this distinction is that in certain situations the two standards can yield opposing prescriptions. This is particularly apparent in the case of hierarchically related categories. By the criterion of coherence, a subordinate category is never more probable than its superordinate. For example, as noted earlier, the probability that someone is British can never be higher than the probability that they are European, because the latter category includes the former. However, it is possible that a particular cue may be more predictive of the subordinate category than of its superordinate. Thus, the information that an individual eats fish and chips (cue C) is a good predictor that he is British (B), but a poor predictor that he is European (E). More formally, even though $P(E|C) > P(B|C)$, $\Delta P_{\mathrm{CB}} > \Delta P_{\mathrm{CE}}$, because $P(B|C) - P(B| \sim C) > P(E|C) - P(E| \sim C)$.

In other words, the presence of cue C raises the probability that the individual is British more than it raises the probability that he is European (indeed it might be thought to lower this latter probability). Nevertheless, the conditional probability that the individual is British, given the cue C, cannot be greater than the conditional probability that he is European, given C. This follows regardless of the content of cue C; it is a formal rule of probability theory.

The crucial point here is that although the predictive cue raises the probability of the subordinate more than it raises the probability of the superordinate, the latter is always, by necessity, at least as high as the former. At the limit, a cue can increase the probability of the subordinate so that it equals that of the superordinate, but it can never raise it higher. More generally, then, whenever a cue is a better predictor of a subordinate category than of the superordinate, there will be a disassociation between predictiveness and coherence. Relative to that cue, the subordinate outcome will be better predicted, but less probable, than its superordinate.

By constructing a hierarchical learning task that sets up this disassociation, it is possible to test a central claim of the associative model. The main requirement is that the environment be structured so that one cue is highly predictive of a target subordinate outcome, but not of its corresponding superordinate category. After exposure to such a learning environment, one can present participants with the predictive cue and ask them to rate the probabilities of both subordinate and superordinate categories. If people do use learned associations as a basis for their probability judgments, rather than an explicit representation of probability, then one would expect them to violate coherence and rate the target subordinate as more probable than its superordinate.

2.2. Overview of experiments

The learning task used in all of our experiments is based on the *medical diagnosis task*,[4] but adapted to include a simple hierarchical category structure. This task has been used to explore a wide range of contingency and category learning phenomena (Estes et al., 1989;

[4] In fact the medical diagnosis task can be viewed as part of a more general experimental paradigm: non-metric multiple cue probability learning (e.g. see Björkman, 1973). The term *non-metric* refers to the discrete-valued cues and category outcomes.

Friedman & Massaro, 1998; Gluck & Bower, 1988; Kruschke & Johansen, 1999; López et al., 1998; Medin & Edelson, 1988; Nosofsky et al., 1992; Shanks, 1990). In its original form, people learn to predict outcomes (diseases) on the basis of *imperfect* indicators (symptoms), and then make judgments on the basis of this learning. There is, therefore, an irreducibly probabilistic relationship between symptoms and diseases, and it is people's knowledge of this structure that is tested in the judgment phase of the task.

The medical diagnosis task thus provides us with a well-tested experimental paradigm. Participants readily understand its rationale, but the stochastic relationship between symptoms and diseases make it quite a challenging task. Furthermore, the medical setting facilitates the introduction of a hierarchical structure, as category hierarchies occur in both expert and lay classifications of diseases. In our adaptation of the task, disease categories are represented at two levels of hierarchy: superordinate (e.g. *Flu*, *Cold*) and subordinate (e.g. *Chinese Flu*, *Russian Flu*).

2.3. Outline of task

The task structure is the same in all of our experiments. In the *training phase* participants undergo a fixed number of learning trials. On each trial, they are presented with a set of symptoms, and must (1) select one of two superordinate disease categories (e.g. *Flu* or *Cold*), then (2) select one of two subordinate categories from that superordinate category (e.g. *Chinese Flu* or *Russian Flu*). Once they have made their selections, feedback is given as to the correct diagnosis (at both category levels). In the *judgment phase* participants are given a sequence of test trials. On each trial a single symptom is presented, and they must estimate the conditional probability of a designated disease (at either the subordinate or superordinate level). The learning environment is constructed so that for some symptom/disease pairs, the symptom is more predictive of the subordinate disease than of its superordinate category. In analyzing participants' actual judgments the critical comparison will be between probability ratings for the subordinate disease and those for the superordinate category.

The associative model predicts that people's probability ratings will be governed by contingency rather than coherence, so that they rate the subordinate greater than the superordinate. In the first two experiments we test this prediction with a standard probability response format. The next three experiments examine ways of reducing incoherence, either with a frequency response format, or an on-screen memory aid.

3. Experiment 1

3.1. Method

3.1.1. Participants and apparatus

Eight students from University College London took part in the experiment. They were paid a fixed sum of £5 for their participation. All participants were tested individually, and the entire experiment was run on a PC computer with a color monitor.

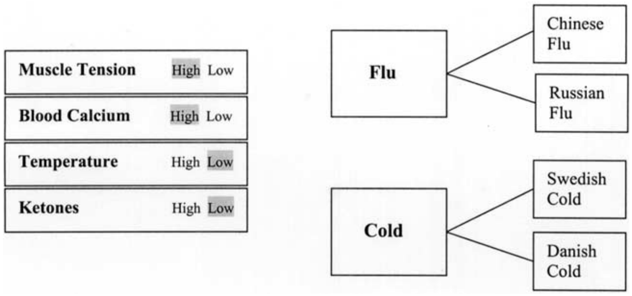

Fig. 2. Screen layout for training phase.

3.1.2. Materials

For the training phase 'medical charts' were constructed for 120 hypothetical patients. Each chart listed the test results, either high or low, for four symptoms: *Muscle Tension*, *Blood Calcium*, *Temperature* and *Ketones*. These were presented on the screen in a vertical list with fixed order. Adjacent to the test results were six disease option boxes, arranged to illustrate the hierarchical nature of the diagnosis problem (see Fig. 2). In the first column there were two option boxes, one labeled with the superordinate disease category 'Flu', the other with the superordinate 'Cold'. In the second column there were four option boxes, each labeled with one of the four subordinate disease categories: 'Chinese Flu', 'Russian Flu', 'Swedish Cold' and 'Danish Cold'.

Each patient chart, which is composed of four possible symptoms $\{s_1, s_2, s_3, s_4\}$, was paired with a single diagnosis, specified at both the superordinate (higher) level $\{h_1, h_2\}$ and subordinate (lower) level $\{l_1, l_2, l_3, l_4\}$. The overall set of symptom/diagnosis pairings made up the participant's learning environment, and is summarized in Table 1.

There were two *target* symptom/diagnosis pairs, $s_1 \rightarrow (l_1\&h_1)$ and $s_4 \rightarrow (l_4\&h_2)$, one for each superordinate category. For each of these pairs:

(1) The conditional probability of the target subordinate disease given the target symptom was set to 0.6, i.e. $P(l_1|s_1) = P(l_4|s_4) = 24/40 = 0.6$, and the conditional probability of the corresponding superordinate given the same symptom was set to 0.7, i.e. $P(h_1|s_1) = P(h_2|s_4) = 28/40 = 0.7$.

(2) The contingency between symptom and subordinate disease was 0.6, whereas the contingency between the same symptom and superordinate was 0.3. That is:

$$\Delta P(l_1, s_1) = P(l_1|s_1) - P(l_1| \sim s_1) = 0.6 - 0 = 0.6$$

$$\Delta P(l_4, s_4) = P(l_4|s_4) - P(l_4| \sim s_4) = 0.6 - 0 = 0.6$$

$$\Delta P(h_1, s_1) = P(h_1|s_1) - P(h_1| \sim s_1) = 0.7 - 0.4 = 0.3$$

$$\Delta P(h_2, s_4) = P(h_2|s_4) - P(h_2| \sim s_4) = 0.7 - 0.4 = 0.3$$

Thus, for each target pairing the symptom was more predictive of the subordinate disease than it was of its superordinate, even though the conditional probability of the superordinate category was greater than that of the subordinate.

Table 1
Trial types in Experiment 1[a]

	h_1		h_2	
	l_1	l_2	l_3	l_4
Target symptoms				
s_1	12	0	0	0
s_1 s_2	6	2	6	0
s_1 s_3	6	2	6	0
s_4	0	0	0	12
s_2 s_4	0	6	2	6
s_3 s_4	0	6	2	6
Non-target symptoms				
s_2	0	5	5	0
s_3	0	5	5	0
s_2 s_3	0	5	5	0
None	0	5	5	0

[a] Target conditional probabilities are as follows: $P(l_1|s_1) = 24/40$; $P(l_4|s_4) = 24/40$; $P(h_1|s_1) = 28/40$; $P(h_2|s_4) = 28/40$.

Once these constraints were set, the finer details of the training set were arranged so as to minimize irrelevant contingencies, and ensure that the task was not unnecessarily difficult for the participants. No individual chart had more than two symptoms reported high, and the two target symptoms (s_1, s_4) were never both high on the same chart. Trials with a target symptom/diagnosis pairing were divided equally between single symptom charts, in which just the target symptom was high, and double symptom charts, in which one additional symptom was also high (either s_2 or s_3). The remaining 40 charts involved non-target pairs. These were equally split between disease l_2 and l_3, with each non-target symptom combination occurring an equal number of times.

For the judgment phase, charts were constructed for a further ten patients. The test results were presented in an identical fashion as in the training phase, but the disease option boxes were replaced with a probability estimate request and a rating slider. The slider was calibrated from 0 (labeled ‘certainly not’) to 100 (labeled ‘certainly’), with the starting position for the pointer at 50 (labeled ‘as likely as not’). Each of the new patients had a single symptom chart, and each chart was paired with a specific diagnosis query. Four of these symptom/diagnosis query pairings were ‘critical’ judgment pairs; the other six were fillers. Each critical judgment pair was made up of a target symptom and a target disease category (either subordinate or superordinate): $s_1 \rightarrow l_1$, $s_1 \rightarrow h_1$, $s_4 \rightarrow l_4$, and $s_4 \rightarrow h_2$. These pairings permitted a direct comparison of the participant’s probability judgments for (1) the target subordinate, given the target symptom, $P(l_1|s_1)$ or $P(l_4|s_4)$, and (2) the superordinate of the target disease, given the target symptom, $P(h_1|s_1)$ or $P(h_2|s_4)$.

3.1.3. Design

In the training phase each participant was exposed to a set of 120 patient charts in random order. Amongst these charts were the two target symptom–disease pairings, differing only in the screen position of the disease categories (either upper or lower). The labeling and screen position of symptoms and diseases were counterbalanced across participants. In the judgment phase each participant gave probability estimates for ten single symptom charts (presented in random order). Included in these were the four critical judgments, one superordinate and one subordinate for each of the two target pairs. Thus, the experiment had a within-subjects design with two factors: hierarchy (superordinate vs. subordinate) and screen position (upper vs. lower).

3.1.4. Procedure

At the start of the experiment participants were told that they would be taking part in a simulated medical diagnosis task. The following instructions were then presented on the screen:

> This experiment examines how people learn to make accurate medical diagnoses. You will be presented with the test results for 120 patients. Some of these are suffering from Flu (either Chinese Flu or Russian Flu), the others from a Cold (either a Swedish Cold or a Danish Cold). For each patient you will be shown a set of test results, and then asked to predict which illness the patient has. After making your diagnosis, you will be shown which illness the patient actually has. All you have to do is try to learn which test results tend to go with which illnesses, so that you can make as many correct diagnoses as possible. Just as with real medical diagnosis, the test results in this experiment are at best only imperfect indicators of the illnesses. You cannot always be correct, but you can learn the tendencies and improve your accuracy. You have as much time as you wish to make your diagnoses. After seeing all 120 patients you will be asked some additional questions about what you have learnt.

Participants were then shown a sample screen layout with annotated instructions. These guided them through the user-controlled procedures for viewing patient charts, registering their diagnoses, and viewing the diagnostic feedback. To view the first chart, participants clicked on a command button labeled ‘Check Test Results’. The test results for each symptom were indicated by a highlighted ‘high’ or ‘low’ adjacent to the symptom label (see Fig. 2). Participants were then required to select one of the two superordinate option boxes. Depending on this choice, they then selected one of the two subordinate diseases. At this stage participants were permitted to change their selections, with the provisos that they select a superordinate prior to a subordinate, and did not select a subordinate from the non-selected superordinate. These constraints were imposed by automatic enabling (or disabling) of the relevant subordinate option boxes. Once they had made their final choices, participants clicked on a ‘Check Diagnosis’ command button. This revealed the correct diagnosis for that chart by highlighting the appropriate superordinate and subordinate disease boxes in red. Participants moved onto the next chart by clicking on ‘Next’. This was repeated for 120 patient charts, presented in a random order for each participant.

At the end of the training phase, participants were shown the following instructions:

Table 2
Mean percentage of correct diagnoses for target pairs

Trial no.	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5
1–30	30	34	32	33	33
31–60	35	40	34	29	53
61–90	49	42	35	39	46
91–120	41	45	40	40	52

> You have now seen all 120 patients. In the final stage of this experiment you will be asked to make some probability judgments on the basis of what you have learnt. You will be presented with the test results for ten new patients. For each new pattern of test results you will be asked to estimate the probability that the patient has a specified illness. To make your estimate you will be provided with a rating scale ranging from 0 to 100. A rating of 0 means that you think the patient is certain not to have the illness in question. A rating of 50 means the patient is as likely as not to have the illness. A rating of 100 means that the patient is certain to have the illness.

Beneath these instructions was a sample rating slider, and participants were invited to familiarize themselves with its use before moving on to the final stage of the experiment. For the judgment phase, participants again had to click on 'Check Test Results' to view the patient chart. The test results were then displayed together with a probability request for one specified disease: 'Given this pattern of test results, what is the probability that the patient has *X*?' They then used the slider to register their judgments, clicking on 'OK' to

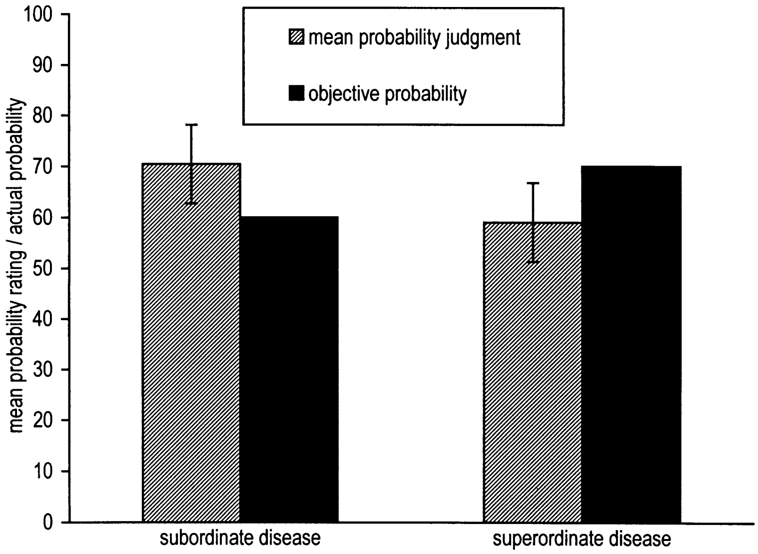

Fig. 3. Mean probability judgments (±within-subject 95% confidence intervals; Loftus & Masson, 1994) and actual conditional probabilities for subordinate and superordinate disease categories in Experiment 1.

move on to the next patient chart. No feedback as to the correct diagnosis was given in this final phase. At the end of the experiment participants were paid and debriefed.

3.2. Results and discussion

Through the training phase participants improved the accuracy of their target diagnoses. Table 2 shows the percentage of correct diagnoses for target pairs across four blocks of 30 trials. Because there were four target diseases, chance performance is 25% correct. As shown in the table, mean correct diagnoses for target pairs increased from 30 to 41%. This suggests that participants had learnt to associate target symptoms with diseases, although their learning was not yet at the optimal level of 60%.

The data of interest are the probability ratings participants gave for subordinate and superordinate diseases, conditional on target symptoms. These are shown in Fig. 3, alongside the actual probabilities occurring in the training set. The ratings combine each participant's ratings for the two target pairs (upper and lower screen position). A two-way analysis of variance (repeated measures) was carried out with hierarchy (subordinate vs. superordinate) and screen position (upper vs. lower) as factors. In line with our predictions, the mean ratings for the target subordinates were higher than for their corresponding superordinates ($F(1,7) = 6.14$, $P < 0.05$). There was no effect of screen position ($F(1,7) = 0.58$), but a marginally significant interaction between hierarchy and screen position ($F(1,7) = 4.66$, $P = 0.07$).[5]

The main result, then, is that the mean rating for the subordinate category is significantly higher than that for the superordinate category, in violation of the rule of extension. This is actually quite a conservative test, because in this experiment (and all of the experiments in this study) the target subordinate is a *proper* subset of its superordinate – i.e. the superordinate category contains other subordinate categories. In such cases the extension rule requires that the superordinate disease be rated *higher* than the subordinate. As shown in Fig. 3, the mean ratings were exactly opposite to that predicted on a normative model. Furthermore, on 14 out of 16 occasions (87%) individual participants made *extension errors*, and rated the subordinate category higher than or equal to the superordinate category. On only two out of 16 occasions (13%) did participants conform to the objective probabilities and rate the superordinate higher than the subordinate; this was significantly less than the nine out of 16 occasions (56%) on which participants rated the subordinate *higher* than the superordinate (sign test, $n = 11$, $P < 0.05$).

4. Experiment 2

This experiment was essentially a replication of Experiment 1 with two minor modifications. First, instead of disease categories such as 'Chinese Flu' or 'Swedish Cold',

[5] This marginally significant interaction reflects the fact that the difference between superordinate and subordinate categories in the upper screen position was larger than for the corresponding categories in the lower screen position. However, in both cases the mean rating for the subordinate was higher than for the superordinate. Furthermore, the comparable interactions (between hierarchy and screen position) in Experiments 2 and 4 do not approach significance.

fictitious disease labels were used. This reduces the possibility of participants' judgments being biased by any prior medical knowledge. Second, the training set was made up entirely of single symptom patient charts. This was designed to simplify the task, allowing participants to focus on the predictiveness of individual symptoms, rather than patterns of symptoms. Furthermore, a single symptom design in the learning phase matches more precisely the single symptom presentations in the judgment phase.

4.1. Method

4.1.1. Participants and apparatus

Twelve students from University College London took part in the experiment. None had participated in the previous experiment. Payment and testing conditions were identical to Experiment 1.

4.1.2. Materials

The stimuli were the same as in Experiment 1 except for two minor changes. First, the disease category labels were changed to fictitious names: one superordinate disease was 'Coralgia', with subordinate categories 'Alpha Coralgia' and 'Beta Coralgia', and the other superordinate was 'Burlosis', with subordinate categories 'Mono Burlosis' and 'Tetra Burlosis'. Second, the training phase was made up entirely of single symptom charts: for each patient chart only one symptom was high. The overall trial structure for the training phase is shown in Table 3. The make-up of charts in the judgment phase was identical to Experiment 1.

4.1.3. Procedure

Procedural details were identical to Experiment 1. Participants diagnosed 120 patients in the training phase, and then made probability judgments for a further ten patients in the judgment phase.

Table 3
Trial types in Experiments 2, 3, 4 and 5[a]

	h_1		h_2	
	l_1	l_2	l_3	l_4
Target symptoms				
s_1	24	4	12	0
s_4	0	12	4	24
Non-target symptoms				
s_2	0	10	10	0
s_3	0	10	10	0

[a] Target conditional probabilities are as follows: $P(l_1|s_1) = 24/40$; $P(l_4|s_4) = 24/40$; $P(h_1|s_1) = 28/40$; $P(h_2|s_4) = 28/40$.

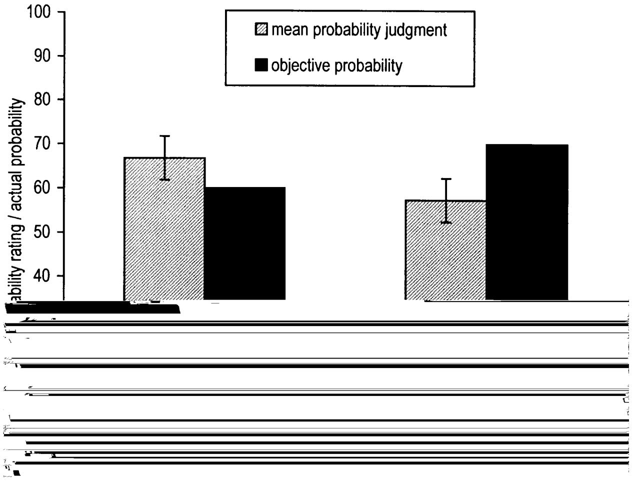

Fig. 4. Mean probability judgments (±within-subject 95% confidence intervals) and actual probabilities for subordinate and superordinate disease categories in Experiment 2.

4.2. Results and discussion

As shown in Table 2, by the end of the training phase participants had improved the accuracy of their target diagnoses from 34 to 45%. As with Experiment 1, this is still below the optimal level of 60%.

The mean probability ratings for subordinate and superordinate categories are shown in Fig. 4, alongside the objective probabilities. Once again, in line with our predictions, the mean ratings for the target subordinates were higher than for their corresponding superordinates ($F(1, 11) = 8.85$, $P < 0.05$). There was no main effect of screen position ($F(1, 11) = 0.09$) and no significant interaction between hierarchy and screen position ($F(1, 11) = 1.75$). Furthermore, in close agreement with Experiment 1, participants committed extension errors on 19 out of 24 occasions (79%), rating the subordinate category either higher than or equal to the superordinate category. On only five out of 24 occasions (21%) did they give normative responses and rate the superordinate higher than the subordinate; this was significantly less than the 14 out of 24 occasions (58%) on which participants rated the subordinate *higher* than the superordinate (sign test, $n = 19$, $P < 0.05$).

Although the results from Experiments 1 and 2 clearly demonstrate that people are incoherent in their probability ratings, and suggest that they are basing their responses on judgments of predictiveness, it might be objected that this is an experimental artifact due to the instructions given to participants prior to the learning phase. In particular, because participants are told to 'learn which test results go with which illnesses', they may interpret their task as one of tracking the predictive relations between cues and outcomes. As a consequence, when asked to give probability judgments in the final stage of the experiment, they may be *primed* to interpret this as a request for estimates of predictiveness.

In a sense this point concurs with our own analysis; as will be discussed below, one of

convert these raw frequencies into probabilities prior to computation, whereas on the frequentist account they use rules that operate directly on the absolute frequencies (e.g. Gigerenzer & Hoffrage, 1995).

In the training phase of our experiments people are presented with information in a sequential fashion. This provides ideal conditions for the natural sampling of event frequencies, and hence the engagement of a frequency-based learning mechanism. To test for an effect of frequency format, we simply asked participants to make frequency rather than probability judgments in the test phase. Quite aside from the speculations about frequency-based mechanisms, it is possible that asking for frequency judgments will alert participants to the fact that the number of patients with a subordinate disease can never exceed the number of patients with the corresponding superordinate disease.

5.1. Method

5.1.1. Participants and apparatus

A further eight students from University College London took part in the experiment. They were paid a fixed sum of £5 for their participation. The same computer set-up was used as in the two previous experiments.

5.1.2. Materials

The training stimuli were identical to those in Experiment 2: fictitious disease labels were used, and the training phase was made up entirely of single symptom charts. The overall trial structure for the training phase is shown in Table 3.

5.1.3. Procedure

Procedural details for the training phase were the same as in the previous experiments. In the test phase, participants were asked to make frequency rather than probability judgments. They were shown the following instructions:

> You have now seen all the patients. In the final stage of this experiment you will be asked to make some frequency judgments on the basis of what you have learnt. You will be presented with another ten patterns of test results. For each you will be asked to imagine that 100 new patients have that pattern of test results. You will then be asked to estimate how many of these 100 patients have a specified illness. To make your estimate you will be provided with a frequency scale ranging from 0 to 100. A rating of 0 means that you think that no patients have the illness in question. A rating of 50 means that half the patients have the illness. A rating of 100 means that all of the patients have the illness.

The subsequent test phase was identical to that in the previous experiments except that each pattern of test results was accompanied by a frequency request: ‘Given that 100 new patients have this pattern of test results, how many have disease *X*?’ As before, participants registered their judgments by using the rating slider marked from 0 to 100.

5.2. Results and discussion

Table 2 shows the percentage of correct target diagnoses in the training phase. Accuracy

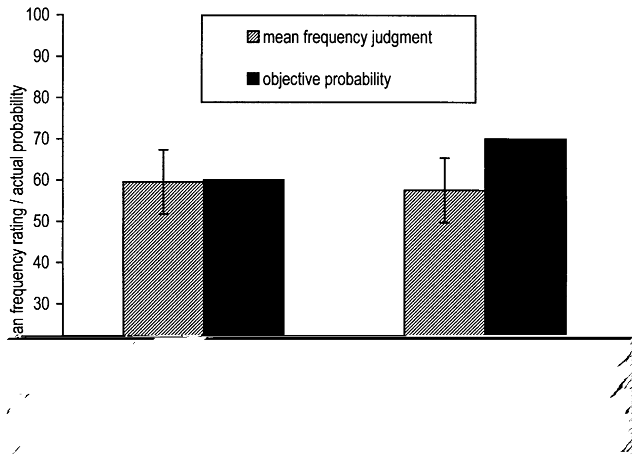

Fig. 5. Mean frequency judgments (±within-subject 95% confidence intervals) and actual probabilities for subordinate and superordinate disease categories in Experiment 3.

improved across blocks from 32 to 40%. The mean frequency ratings for subordinate and superordinate categories are shown in Fig. 5, alongside the objective frequencies. There was no significant difference between subordinate and superordinate ratings ($F(1,7) = 0.18$). There was also no effect of screen position ($F(1,7) = 0.45$) nor an interaction between hierarchy and screen position ($F(1,7) = 2.32$).

Although there was no difference in the mean ratings for subordinate and superordinate, extension errors were still committed on 11 out of 16 (69%) occasions. However, the number of normative responses (five out of 16, 31%) was the same as the number of occasions on which participants rated the subordinate *higher* than the superordinate.

In sum, using a frequency rather than a probability response format appears to somewhat reduce the extent to which people commit extension errors, and push their ratings in the normative direction. To replicate this finding on a larger scale, a follow-up experiment was designed with response format (probability or frequency) as a between-subject factor.

6. Experiment 4: a replication of the frequency format effect

The results from Experiments 1, 2 and 3 suggest that a frequency format tends to reduce the counter-normative nature of people's ratings relative to a probability format. This experiment aims to replicate this pattern within a single experiment.

6.1. Method

6.1.1. Participants and apparatus

Twenty-four students from University College London took part in the experiment. All were naive to the task, and payment and testing conditions were the same as in previous experiments.

individual's ratings diverge from normativity – according to the objective probabilities both difference scores should equal $70 - 60 = 10$. Indeed, any non-positive score is counter-normative. In Fig. 7 each point represents the two difference scores from each participant. It is clear that participants in the frequency condition were more normative than those in the probability condition – the individual points tend to be higher and further to the right.

It is worth noting that although the frequency format clearly improves the coherence of people's ratings, a sizable proportion of them (69% in Experiment 3, 62% in Experiment 4) still violated the extension rule. This suggests that whatever the processes or mechanisms that serve to improve performance under the frequency format, there remains a strong pull towards responses based on predictiveness. This issue will be explored further in Section 8.

7. Experiment 5

This experiment was designed to investigate whether presenting participants with summary frequency information during the training phase reduced susceptibility to extension errors in the test phase. The provision of such frequency information, updated on each trial, eliminates the need for participants to learn associations during the prediction task. Without the incentive for predictive learning, would participants still rate the target subordinate category higher than its superordinate?

7.1. Method

7.1.1. Participants and apparatus

Eight students from University College London took part in the experiment. None had participated in any previous experiments. Payment and testing conditions were the same as in the previous experiments.

7.1.2. Materials

The training and test stimuli were identical to those used in Experiment 2. The only difference was in the display presented to participants during the training phase. This included a *previous cases box* (see Fig. 8), which contained frequency information about the past occurrences of diseases given the currently presented symptom pattern. The box was positioned underneath the symptom list, on the left side of the screen.

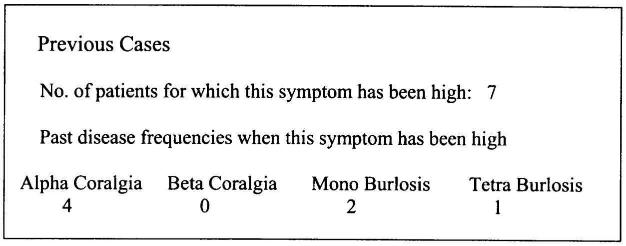

Fig. 8. Previous cases box displayed and updated during the training phase in Experiment 5.

7.1.3. Procedure

Procedural details were the same as in the previous experiments, except participants were instructed to use the previous history information as an aid to diagnosis. The previous cases box was updated after each trial, and showed the relevant frequencies for the currently presented symptom pattern. Recall that each pattern comprised only a single symptom with a high value.

7.2. Results and discussion

The percentage of correct target diagnoses during the training phase is shown in Table 2. Accuracy improved across blocks from 33 to 52%. This may not reflect superior learning of the associations between symptoms and diseases, however, because after the first block the frequency information alone was a sufficient basis for optimal diagnosis.

The mean probability ratings for subordinate and superordinate categories are shown in Fig. 9, alongside the objective probabilities. Although the mean rating for the superordinate category (65.8) was higher than for the subordinate category (57.2), this was not quite significant ($F(1, 7) = 3.68$, $P = 0.097$). There was no effect of screen position ($F(1, 7) = 0.35$) and no interaction between hierarchy and screen position ($F(1, 7) = 0.04$). Extension errors were reduced to eight out of 16 (50%), and on eight out of 16 (50%) occasions responses were in the normatively correct direction.

This experiment served as a kind of control – when participants could use frequency information during training, and thus bypass the need to learn associations between symptoms and diseases, they made judgments in line with the normative probabilities.

Furthermore, it also shows that the extension effect is not simply due to a preference for the more precise category, irrespective of memory and learning. When participants do not need to rely on the latter processes their judgments approach normativity.

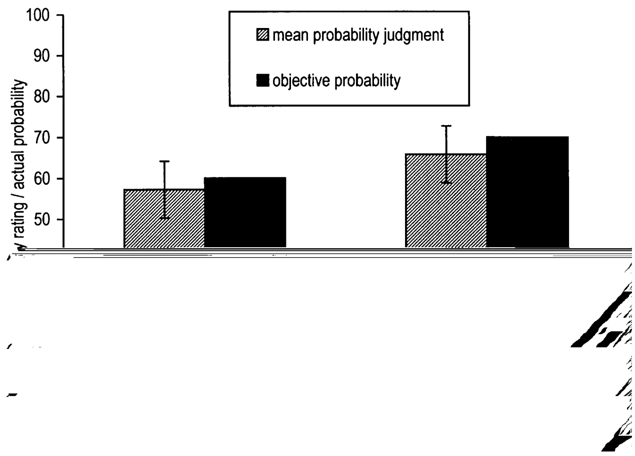

Fig. 9. Mean probability judgments (±within-subject 95% confidence intervals) and actual probabilities for subordinate and superordinate disease categories in Experiment 5.

judgment. Before discussing a dual-process model, however, it is worth considering how the associative approach might be extended without compromising its status as a single-system account.

8.2.1. An associative-plus-adjustment model

One way for the frequency format effect to be accommodated within an associative framework, without admitting a separate frequency-encoding component, is to introduce some kind of *adjustment* mechanism at the judgment phase. Thus, learned associative strengths still remain the primary source for judgments, but certain features of the response format, i.e. frequency rather than probability frames, may lead to the adjustment of this raw value. In particular, a frequency response format may alert people to the set inclusion relations amongst the possible outcomes, and make them adjust their estimates appropriately.

In our experiments, then, the frequency response format may have alerted participants to the fact that items in a subordinate category cannot be more numerous than those in the superordinate category. Without recourse to a 'true' estimate of the actual frequencies, people still use an associative strength, but adjust this value upwards or downwards to obey the set inclusion relation.

Despite the ad hoc nature of this speculation, the consistent overrating of the subordinate categories observed in Experiment 4, under both probability and frequency formats, lends it a little support. The strong predictive link between cue and subordinate outcome may have led to an overestimation of the subordinate category, regardless of rating format. In contrast, the superordinate category is underrated in the probability condition, in line with its weaker link with the cue, but adjusted upwards as a result of the frequency condition. One problem with this suggestion lies in the difficulty in generalizing across the two response conditions. Can one assume that the mapping from associative strength to an overt estimate is the same in both conditions? Further studies looking at within-subject manipulations of response format would better test out this claim.

This approach to explaining the frequency format effect is very similar to a proposal made by Kahneman and Tversky (1996) in their discussion of the conjunction fallacy. They argue that a frequency format can facilitate performance in the conjunction problem because it provides extensional cues to set inclusion relations. The basic claim is that under both probability and frequency formats people use the representativeness heuristic to reach an estimate, but if the problem format provides the appropriate extensional cues they adjust their estimates accordingly: "subjects use representativeness to estimate outcome frequencies and edit their responses to obey class inclusion in the presence of strong extensional cues" (Kahneman & Tversky, 1996, p. 587).[6]

The associative-plus-adjustment model improves on Kahneman and Tversky's suggestion in so far as it replaces the vague notion of representativeness with the more precise one of predictiveness. But it shares with it the shortcomings of an imprecise and under-specified heuristic of adjustment or editing.

[6] This kind of position has been defended recently by Sloman and Over (in press) and Girotto and Gonzalez (2001).

8.2.2. A dual-component model

The most natural interpretation of the findings from our experiments is that people possess both associative and frequency-encoding mechanisms. Both of these are presumed to operate during learning, but the output of one may take precedence over the other in the judgment phase. This will depend on the nature of the judgment task. If the request is for the probability of outcome given cue, people tend to base their judgment on the learned association between cue and outcome – thus conflating probability with predictiveness, and possibly violating the rule of extension. If the request is for the relative frequency of outcomes, people are more likely to base their judgment on the appropriate encoded frequencies – and thus conform to the extension rule.

In their studies of contingency learning and judgment, Price and Yates (1995) also advanced a dual-component model. In part this was motivated by frequency effects akin to those found in our experiments. Their model draws upon a distinction made by Hastie and Park (1986) between *on-line* and *memory-based* judgment tasks. In an on-line task, people are encouraged or required to integrate the stimulus data as they are presented, whereas in a memory-based task they are encouraged to store data in a raw form, and integrate it only when requested to make a judgment. Most predictive learning tasks, including the diagnosis task used in our experiments, would fall into the former category. Price and Yates maintain that these on-line tasks are likely to engage people in trial-by-trial associative learning. Irrespective of whether the task is on-line, however, people are also supposed to store event frequencies. When required to make a contingency judgment, people either base this on an appropriate association strength, or, in the absence of this, compute it on the basis of the encoded frequencies.

Although primarily concerned with contingency judgments, Price and Yates' model is readily adapted to cover judgments of probability and frequency. In this case the nature of the judgment request, i.e. whether it is framed in terms of probabilities or frequencies, would be expected to determine whether it is based on a learned associative strength or on encoded frequencies. Furthermore, the demand for a conditional probability – the probability of outcome *given* cue – might increase the bias towards the reporting of an associative strength, because it suggests a concern with the strength of relation between cue and outcome.

Thus, a dual-component model, along the lines suggested by Price and Yates, readily accommodates the data from our experiments. Several additional factors also recommend such a model. For one, there is a significant body of evidence that suggests that humans are reasonably accurate at monitoring event frequencies (Barsalou & Ross, 1986; Estes, 1976; Hasher & Zacks, 1984; Jonides & Jones, 1992). It is not clear how these findings could be explicated without postulating some kind of frequency-encoding mechanism. A related point is how one could explain accurate probability (or frequency judgments) when there are no predictive relationships between the events to be assessed. After exposure to a zero contingency, participants would be rather unlikely to rate the conditional probability as zero. Presumably, in the absence of contingency information, people base their judgments on base rates derived from stored frequencies. Third, philosophers and psychologists have often noted the ambiguous nature of the term 'probability' in everyday discourse (e.g. Carnap, 1950; Hacking, 1975, 2001; Hertwig & Gigerenzer, 1999), and have distinguished between two main interpretations: (1) an *epistemic* or belief-based concept (e.g. as a relation between evidence and hypothesis); (2) a frequency-based concept. Indeed, Carnap (1950) suggested that the terms probability$_1$ and probability$_2$ be introduced to mark this distinction. Perhaps it is not too surprising that such a demarcation is mirrored in the cognitive system that underlies people's actual probability judgments.[7] Thus, when asked for the probability of outcome given cue, people may well be orientated towards (1), and base their judgment on the strength of association between cue (evidence) and outcome (hypothesis). However, a frequency request may orientate them to (2), and a judgment based on encoded frequencies.

To summarize, then, the data in our experiments, in particular the extension and frequency effects, seem to be best accounted for within a dual-component model. Whilst there is no significant loss in parsimony, since any unitary model would also require supplementary mechanisms, there is a substantial gain in the range of data to which it applies.

8.3. Relation to other research

Most of the research on probability judgment has been conducted in *described* rather than *experienced* paradigms. Although one would expect strong commonalities across the two domains, it is also possible that they engage quite different kinds of cognitive mechanism. With this proviso in mind, we will apply some of the findings and arguments from this paper to certain issues in the judgment and decision literature.

8.3.1. The conjunction fallacy

The conjunction fallacy is closely related to the extension error studied in this paper. It too involves a violation of the extension rule of probability, but the focus is on the subgroup formed by the conjunction of two different superordinate groups. The conjunction fallacy arises when this subgroup is rated as more probable than one of its superordinate groups. The classic demonstration of this fallacy is given by the Linda problem (Tversky & Kahneman, 1983). Participants are presented with a brief profile of Linda. This lists properties that are supposed to be representative of a feminist but unrepresentative of a bank teller. Participants are then required to rank (or in later variations to rate) her likely membership of various categories, including *feminist* (F), *bank teller* (B), and *bank teller & feminist* ($B\&F$). The major finding, replicated innumerable times and with a variety of cover stories, is that people give a rank ordering of $F > B\&F > B$. This violates the rule of extension, because the probability of the conjunction *bank teller & feminist* can never be greater than the probability of one of its conjuncts *bank teller*, i.e. $P(B\&F|L) \leq P(B|L)$. Tversky and Kahneman explain this result in terms of the representativeness heuristic – people do not base their intuitive judgments on an extensional notion of probability, but instead rely on a judgment about how well Linda's profile *represents* the various cate-

[7] In a sense a dual-component model provides a process-level account for how people disambiguate the term 'probability' in a given learning context. Prior performance in an on-line learning task will engage associative mechanisms, and lead people to interpret a conditional probability request as a question about the relation between evidence and outcome. Once orientated towards this reading of the term people base their judgments on the degree to which the evidence predicts the outcome, rather than on the experienced frequencies.

gories. They rank *bank teller* & *feminist* as more probable than *bank teller* because the profile is more representative of the former than the latter.

It is impossible to summarize the wealth of responses that this problem has evoked. One major line of criticism is voiced by proponents of the frequentist school (e.g. Cosmides & Tooby, 1996; Gigerenzer, 2000). They argue that such problems have been presented in an inappropriate format – one ill-suited to the mind's natural design. According to this view, it is natural frequencies, not percentages, proportions, or probabilities, that the mind is adapted to process. In support of this conjecture, numerous studies have shown that once an appropriate frequency format is adopted, the conjunction fallacy is significantly reduced (e.g. Fiedler, 1988; Hertwig & Gigerenzer, 1999).

Although the introduction of a frequency problem representation does successfully reduce the conjunction effect, and is akin to the frequency format effect observed in our data, this line of reasoning is not entirely satisfactory. One objection, presented in a recent study by Sloman and Over (in press), is that it may not be the frequency format per se that reduces errors, but the cues to set inclusion which the frequency format provides. More pertinent for our present purposes, the frequentist approach does not explain why people make the errors in the first place under a probability format. Neither Gigerenzer (1991, 2000) nor Sloman and Over (in press) address this problem directly.

We believe that the analysis provided in this paper suggests an answer to this question, and in particular, that the propensity for people to conflate probability with predictiveness that leads to extension errors in our experiments may likewise be responsible for conjunction errors.

For one, the standard presentation of the conjunction problem strongly encourages people to assess the extent to which the profile makes the target category probable. Moreover, as argued in this paper, it is natural for people to interpret this as a request for the degree to which the profile *predicts* the category. But in the standard problem Linda's profile (L) is more predictive of the conjunction *bank teller* & *feminist* than just *bank teller*. That is: $\Delta P(L \rightarrow B\&F) > \Delta P(L \rightarrow B)$ because $P(B\&F|L) - P(B\&F| \sim L) > P(B|L) - P(B| \sim L)$.

In other words, the information given by Linda's profile raises the probability of her being a *bank teller* & *feminist* more than it raises the probability of her being a *bank teller* – in fact it may be judged to lower this latter probability. Hence, if people do conflate probability with predictiveness, then they will rate the probability of *bank teller* & *feminist* higher than the probability of *bank teller*, and commit a conjunction error.

This kind of explanation actually fits quite well with a recent analysis of the conjunction problem given by Hertwig and Gigerenzer (1999). They argue that conjunction errors in the probability format are due to participants inferring various non-mathematical senses of probability. In effect our suggestion amounts to replacing this rather vague bundle of non-mathematical interpretations of probability with the more precise notion of predictiveness, a notion that can be objectively measured and experimentally manipulated.

It is important to note that our main claim here is that people often utilize predictive or associative knowledge to reach their probability judgments. It is not essential, however, that they acquire this knowledge through incremental associative learning. Thus, the analysis of probabilistic incoherence offered in this paper is still applicable to the 'one-shot' problems typical of the judgment and decision research. In short, in described problems such as the conjunction and disjunction tasks (Bar-Hillel & Neter, 1993), people may still base their judgments on associative knowledge, and thus conflate probability with predictiveness, even though this knowledge has not been arrived at through trial-by-trial associative learning.

8.3.2. A trade-off between informativeness and accuracy

So far in this paper the rating of subordinate over its superordinate has been characterized as an extension *error*, and by implication as counter-normative. However, there is a sense in which such an error, when translated into a preference for *predicting* the subordinate rather than the superordinate category, is justified pragmatically and hence rationally. The basic intuition here is that predictive inference should not be appraised solely with respect to its coherence with the laws of probability, nor simply with respect to the likelihood that it is correct. A further dimension for appraisal, and one that is actually inversely proportional to probability, is the *informativeness* of the inference. Thus, a more specific or precise inference – such as the subordinate category *Asian flu* – conveys more information than its superordinate category *Flu*, although it is correspondingly less probable.

One proposal gaining currency in the judgment and decision literature is that when making or appraising numerical estimates people negotiate a trade-off between the accuracy of an inference (the probability that it is correct) and the possible gain in informativeness (Yaniv & Foster, 1995, 1997).[8] Applied to our experiments, the claim would be that in making an extension error, and preferring the subordinate to the superordinate category, participants are in fact trading a small probability of being correct for a substantial gain in precision. Indeed, such behavior, when placed in the context of everyday or expert reasoning, seems both pervasive and pragmatic. For example, a doctor may prefer to make a more specific diagnosis, especially if this leads to a more definite treatment, even though it entails a slightly higher likelihood of being incorrect.

So, a more generous reading of the extension errors observed in our experiments recasts them as a trade-off between the probability of being correct and a potential gain in precision. This reading rests on two suppositions. First, that people's probability judgments reflect their preferred inferences. That is, that rating the subordinate as more probable than the superordinate in some way incorporates their preference to infer the former over the latter. Future studies could test this assumption more directly, by getting participants to choose between subordinate and superordinate options – indeed, this is what Bar-Hillel and Neter (1993) did in their experiments, and found that people consistently preferred the subordinate inferences, even in monetary bets. A second supposition is that people compute this trade-off implicitly rather than explicitly. That is, when making or assessing a categorical inference, they do not form an explicit representation of its probability, and of its informational content, and somehow balance these two values. Rather, by following the dictates of predictiveness, they make an inference that trades off these factors automatically.

[8] A similar trade-off has been discussed in the categorization literature (Corter & Gluck, 1992; Murphy & Lassaline, 1997; Rosch et al., 1976). A central claim is that the basic level provides an optimal balance between accuracy of categorization and informativeness (measured in terms of predictive power).

Acknowledgements

This research was supported by grants from the United Kingdom Economic and Social Research Council (ESRC) and the Leverhulme Trust. The work is part of the programme of the ESRC Centre for Economic Learning and Social Evolution, University College London. We thank two anonymous reviewers for their helpful comments.

Appendix A. Probability explicit instructions for replication of Experiment 2

A.1. Instructions for learning phase

This experiment examines how people make accurate probability judgments. You will be presented with the test results for 120 patients. Some of these are suffering from Burlosis (either Mono Burlosis or Tetra Burlosis), the others from Coralgia (either Alpha Coralgia or Beta Coralgia).

For each patient you will be shown a set of test results, and then asked to make a diagnosis. After making your diagnosis, you will be shown which disease the patient actually has.

All you have to do is try to learn the probabilities of the diseases, conditional on the different patterns of test results.

You have as much time as you wish to make your diagnoses. After seeing all 120 patients you will be asked to make some conditional probability judgments about the diseases.

A.2. Instructions for judgment phase

You have now seen all 120 patients. In the final stage of this experiment you will be asked to make some probability judgments.

You will be presented with the test results for ten new patients. For each new pattern of test results you will be asked to estimate the probability that the patient has a specified illness.

To make your estimate you will be provided with a rating scale ranging from 0 to 100. A rating of 0 means that you think the patient is certain not to have the illness in question. A rating of 50 means the patient is as likely as not to have the illness. A rating of 100 means that the patient is certain to have the illness.

References

Allan, L. G. (1993). Human contingency judgments: rule-based or associative? *Psychological Bulletin*, *114*, 435–448.

Allan, L. G., & Jenkins, H. M. (1980). The judgment of contingency and the nature of the response alternatives. *Canadian Journal of Psychology*, *34*, 1–11.

Bar-Hillel, M., & Neter, E. (1993). How alike it is versus how likely it is: a disjunction fallacy in probability judgments. *Journal of Personality and Social Psychology*, *65*, 1119–1131.

Barsalou, L. W., & Ross, B. H. (1986). The roles of automatic and strategic processing in sensitivity to superordinate and property frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*, 116–134.

Björkman, M. (1973). Inference behavior in nonmetric ecologies. In L. Rappoport & D. A. Summers (Eds.), *Human judgment and social interaction*. New York: Holt, Rinehart & Winston.

Carnap, R. (1950). *Logical foundations of probability*. Chicago, IL: University of Chicago Press.

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

López, F. J., Cobos, P. L., Caño, A., & Shanks, D. R. (1998). The rational analysis of human causal and probability judgment. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 314–352). Oxford: Oxford University Press.

Medin, D. L. (1983). Structural principles in categorization. In T. Tighe & B. Shepp (Eds.), *Perception, cognition and development: interactional analyses*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, *117*, 68–85.

Murphy, G. L. (1982). Cue validity and basic levels in categorization. *Psychological Bulletin*, *91*, 174–177.

Murphy, G. L., & Lassaline, M. E. (1997). Hierarchical structure in concepts and the basic level of categorization. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts and categories*. Hove: Psychology Press.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 211–233.

Osherson, D. N. (1995). Probability judgment. In E. E. Smith & D. N. Osherson (Eds.), *Thinking* (pp. 35–75), 3. Cambridge, MA: MIT Press.

Price, P. C., & Yates, J. F. (1995). Associative and rule-based accounts of cue interaction in contingency judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21* (6), 1639–1655.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Current theory and research-Classical conditioning* (pp. 64–99), II. New York: Appleton-Century-Crofts.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Shanks, D. R. (1990). Connectionism and the learning of probabilistic concepts. *Quarterly Journal of Experimental Psychology*, *42A*, 209–237.

Shanks, D. R. (1991). A connectionist account of base-rate biases in categorization. *Connection Science*, *3*, 143–162.

Shanks, D. R. (1995). *The psychology of associative learning*. Cambridge: Cambridge University Press.

Shanks, D. R., López, F. J., Darby, R. J., & Dickinson, A. (1996). Distinguishing associative and probabilistic contrast theories of human contingency judgment. In D. R. Shanks, K. J. Holyoak & D. L. Medin (Eds.), *The psychology of learning and motivation: causal learning* (pp. 265–311), 34. San Diego, CA: Academic Press.

Sloman, S. A., & Over, D. E. (in press). Probability judgment: from the inside and out. In D. E. Over (Ed.), *Evolution and the psychology of thinking: the debate*. Hove, UK: Psychology Press.

Stone, G. O. (1986). An analysis of the delta rule and the learning of statistical associations. In D. E. Rumelhart, J. L. McClelland & the PDP Research Group (Eds.), *Parallel distributed processing: explorations in the microstructure of cognition* (pp. 444–459), 1. Cambridge, MA: MIT Press.

Tversky, A., & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: heuristics and biases*. Cambridge: Cambridge University Press.