

Running head: COHERENT PROBABILITY FROM INCOHERENT JUDGMENT

Coherent probability from incoherent judgment

Daniel Osherson David Lane Peter Hartley Richard R. Batsell

August 8, 2000

Abstract

People often have knowledge about the chances of events but are unable to express the knowledge in the form of coherent probabilities. We propose to correct incoherent judgment via an optimization procedure that seeks the (coherent) probability distribution nearest to the judge's estimates of chance. Our method was applied to the chances of simple and complex meteorological events, as estimated by college undergraduates. No judge responded coherently but our optimization method found close (coherent) approximations to their estimates. Moreover, the approximations were reliably more accurate than the original estimates, as measured by the quadratic scoring rule. Methods for correcting incoherence facilitate the analysis of expected utility, and allow human judgment to be more easily exploited in the construction of expert systems.

Coherent probability from incoherent judgment

Introduction

Suppose you think the probability that the internet will expand next year is .90. Suppose you also think the probability that the internet will expand and PC makers will be profitable is .91. Then you have assigned a greater chance to a conjunction than to one of its conjuncts, hence your judgments are incoherent. You may nonetheless prove to be more insightful than someone with the coherent opinion that internet expansion next year has probability .2 and that both expansion and PC profitability has probability .1. This example indicates that judgments may be rich in information without being probabilistically coherent. This is reassuring since it is well known that maintaining coherence is a daunting task both for computers and for human judges.¹

Three responses to incoherent estimates of chance may be envisioned. They are:

- (a) Do nothing. Live with incoherence.
- (b) Prevent incoherent judgment through a structured elicitation technique.
- (c) Repair incoherence after-the-fact by adjusting numerical estimates of chance.

The first response is risky since incoherent judgments lead to systematic losses (“Dutch Books”) when spotted by an adversary (provided the judge is willing to accept bets s/he deems fair). Specifically, it has been proven that a set of judgments is incoherent if and only if there are monetary bets with the following properties: (a) each bet has zero expected monetary value according to the judgments, i.e., each bet seems fair to the judge; but (b) the net outcome of the bets for the judge is negative no matter which events in the bets occur.² Incoherent assessments of chance are also inimical to the analysis of expected utility, which relies on genuine estimates of probability (see Kleindorfer, Kunreuther and Schoemaker, 1993; for the role of coherence in standard justifications of utility theory, see Jeffrey, 1983, Chapter 4). In addition, incoherent judgment cannot be incorporated into Bayesian networks, one of the most popular approaches to automated reasoning (Castillo, Gutiérrez and Hadi, 1998). Methods for avoiding or correcting incoherent judgments of probability therefore represent a potential contribution to individual decision-making and to other disciplines (e.g., economics, computer science).

Approach (b) stamps out incoherence at the source by monitoring the successive estimates emitted by the judge. Probabilities of a given event are required to be drawn from an interval of coherent possibilities, calculated from earlier judgments. Variations on this idea have produced important innovations in elicitation methods.³ Eliminating incoherence during elicitation can be a tedious procedure, however, and may alter the judge’s opinions.

Some arbitrariness is also to be expected, since the final set of estimates will likely depend on the order in which judgments are elicited. Structured elicitation, moreover, does not apply to situations in which judges are polled by written questionnaire, or brief interview. It would be desirable to have a method for adjusting estimates of chance after-the-fact, bringing them into coherence even if the judge has left the scene. Such an off-line approach to coherence is envisioned in (c), and explored in the present paper.

So far as we know, the first schemes of kind (c) were described in Lindley, Tversky and Brown (1979). These authors conceive the judge's estimates as arising via error from an underlying source of coherent probabilities (not consciously accessible to the judge herself). The task of an observer is to infer the coherent probabilities from the incoherent stated ones. This is achieved on the basis of the observer's prior distribution over the potential coherent beliefs the judge might secretly harbor, along with another prior distribution that gives the probability of stated beliefs given (coherent) underlying ones. An application of Bayes' theorem then allows calculation of the most likely underlying assessments of chance given the stated ones. Various simplifying assumptions allow the desired calculations to be formulated perspicuously, but reaching a solution ultimately requires nonlinear optimization. A second approach described in the same paper relies on a similar set of prior distributions, exploited somewhat differently.

Lindley et al.'s (1979) approach is pathbreaking and ingenious, but it requires specifying prior distributions that are difficult to interpret and to evaluate, a point raised by the authors themselves as well as by commentators on the paper. Since the specification of these distributions does not obviate a complex optimization step, it strikes us as simpler to forgo the former and proceed at once to the latter. We shall therefore conceive of off-line correction as the search for a (coherent) probability distribution that best approximates the incoherent probabilities in hand.

An algorithm for finding coherent approximations will be proposed below. It rests on the following hypothesis about the probability judgment of well-informed human agents.

- (i) Hypothesis of sparse distributions: If the agent's judgments can be approximated by any probability distribution, they can be approximated by a distribution that assigns positive probability to a relatively small number of potential states-of-affairs.

To illustrate, suppose that an agent is considering the ranks of ten teams at the conclusion of a tournament. There are millions of possible orderings but Hypothesis (i) predicts that the agent's judgments can be approximated by a distribution that assigns positive probability to only a few of them. The remainder are "edited" out of consideration via the assignment of zero probability. The hypothesis is based on the observation that human reasoners seem to hold only a few alternative possibilities in mind (see Manktelow, 1999, chapter 10, for discussion). It receives support from the generally successful performance of our algorithm. Note that the hypothesis is intended to apply only to knowledgeable agents. Ignorance might lead to the adoption of the uniform distribution, which is not sparse.

As a final preliminary, we note that only point probabilities are at issue in the present discussion. Point probabilities are typically elicited in the construction of decision trees and influence diagrams (as in the popular program DATA 3.5 by TreeAge Software). Moreover, allowing judges to offer probability intervals or distributions over probabilities introduces complexities that are best set aside in a preliminary study like this one. (For analysis of imprecise estimates of chance, see Walley, 1991, 1996).

To proceed, we first define the concept of probabilistic coherence. Next we describe our method for calculating a (coherent) distribution that approximates an input set of probability estimates. We then turn to an experimental test of the method, involving the collection of probabilistic weather forecasts from undergraduates. It will be seen that the judgments were indeed incoherent, but well approximated by the optimization method we employ. Moreover, the coherent approximations provide reliably more accurate forecasts than the original estimates.

Theoretical background

Probability

The conception of probability summarized in the present section is due to de Finetti (1972). Nilsson (1986) provides more ample discussion than possible here. In overview, we let 2^n states be generated by n binary variables. The states are assigned probabilities, which are extended to events (sets of states) in the usual way. Formulas of sentential logic are used to name events, and inherit the latter's probabilities. The formulas are translated into event-descriptions of English. Conditional probability is handled in parallel manner. We now provide details.

Assume that all the events under consideration can be represented as boolean combinations of n variables. Each variable takes one of the two truth values true (t) and false (f). By a state is meant any map of the n variables into $\{t, f\}$. A state is thus a potential "state-of-affairs" which determines the truth-value of every variable. The n variables yield 2^n states. By a (probability) distribution for the n variables is meant any mapping Pr of the states into $[0, 1]$ such that $\sum\{\text{Pr}(s) : s \text{ is a state}\} = 1$. To illustrate, suppose that there are just three variables, p, q, r . Then one distribution over the 8 resulting states is as follows.

state	p	q	r	Pr
(1)	t	t	t	.15
(2)	t	t	f	.15
(3)	t	f	t	.10
(4)	t	f	f	.10
(5)	f	t	t	.10
(6)	f	t	f	.10
(7)	f	f	t	.15
(8)	f	f	f	.15

Each row in (ii) corresponds to a state, e.g., the one in which all three of p, q, r are true. There are 8 states since each of the 3 variables can independently assume either truth-value. The last column of (ii) represents a distribution because it associates a nonnegative number with each state in such a way that the numbers sum to unity.

We now consider how a given distribution imposes probabilities on statements. The statements in question are described by the kind of formal language familiar from sentential logic. Specifically, our language includes the n variables as formulas, and is then built up using the sentential connectives in the usual way. Thus, it consists of negations (e.g., $\neg p$), conjunctions (e.g., $p \wedge r$ and $p \wedge \neg q$), disjunctions (e.g., $r \vee q$ and $(r \wedge q) \vee \neg p$), among other types of formulas. We presuppose the concept (familiar from sentential logic) that a given state makes true a given formula. For example, states 1 - 4 make true p , states 2, 6 make true $q \wedge \neg r$, and states 1, 2, 5 - 8 make true $\neg p \vee q$. A given formula represents the event consisting of the states that make it true. Intuitively, to assert formula φ is to claim that one of the states making φ true is the actual state-of-affairs. Thus, to assert p is to claim that the world conforms to one of the states 1 - 4.

Since states are mutually exclusive, and each formula represents a set of them, it is clear how to extend a given distribution \Pr to the formulas of our sentential language. For every formula φ ,

$$\Pr(\varphi) = \sum \{\Pr(s) : s \text{ is a state that makes } \varphi \text{ true}\}.$$

That is, the probability of a formula is the sum of the probabilities of the states that make it true. For example, if \Pr is shown in (ii), then $\Pr(p) = .15 + .15 + .10 + .10 = .5$, $\Pr(q \wedge \neg r) = .15 + .10 = .25$, and $\Pr(\neg p \vee q) = .10 + .10 + .15 + .15 + .15 = .8$.

For conditional probabilities, \Pr is extended again, this time to pairs of formulas. The pair of formulas consisting of φ followed by ψ is standardly written (φ, ψ) . When writing conditional probabilities, however, it is customary to use the symbol $|$ in place of a comma to separate two formulas in a pair. Thus, when the probability function is applied, the foregoing pair of formulas is written $(\varphi | \psi)$. The pair can be read “ φ assuming that ψ .” This translation is revealing of the intended interpretation of pairs of formulas but there is an important caveat. Although “assuming that” behaves like a sentential connective in English, it is well known that the symbol $|$ cannot be interpreted as a sentential connective analogously to \wedge or \vee (Lewis, 1976; Bradley, 1999). We can now say what number a given distribution \Pr assigns to a pair of formulas. For formulas φ, ψ with $\Pr(\psi) > 0$,

$$\Pr(\varphi | \psi) = \frac{\sum \{\Pr(s) : s \text{ is a state that makes both } \varphi \text{ and } \psi \text{ true}\}}{\sum \{\Pr(s) : s \text{ is a state that makes } \psi \text{ true}\}}.$$

For example, $\Pr(q \wedge \neg r | p) = \frac{.15}{.15 + .15 + .10 + .10} = .3$. Since a state makes a conjunction true just in case it makes true both conjuncts, the definition of conditional probability implies the familiar fact that for formulas φ, ψ with $\Pr(\psi) > 0$,

$$\Pr(\varphi | \psi) = \frac{\Pr(\varphi \wedge \psi)}{\Pr(\psi)}.$$

Observe that \Pr makes no assignment of probability to a pair (φ, ψ) if it assigns zero probability (impossibility) to the conditioning event ψ .

Coherence

Consider a judge who is estimating the probabilities of various events represented in our sentential language. We write $Prob(\varphi) = x$ to indicate the judgment that the probability of φ is x , and $Prob(\varphi \mid \psi) = y$ for the judgment that the conditional probability of φ assuming ψ is y . It is important to distinguish $Prob$ from the kind of function denoted by \Pr . $Prob$ is no more than a mapping of some formulas and pairs of formulas into numbers. Its domain will represent whatever (possibly disparate) collection of statements were evaluated by the judge, and hence be finite (unlike the domain of \Pr , which embraces every formula of our sentential language). In particular, $Prob$ need not conform to any of the properties that apply to genuine probability distributions like \Pr . This is why the judge's estimates are not written with the symbol \Pr , which is reserved for genuine distributions. To illustrate, $Prob$ might be the following set of judgments.

$$(iii) \quad \begin{aligned} Prob(p) &= .5 \\ Prob(q \wedge \neg r) &= .25 \\ Prob(\neg p \vee q) &= .8 \\ Prob(q \wedge \neg r \mid p) &= .3. \end{aligned}$$

We call $Prob$ probabilistically coherent just in case there is a probability distribution \Pr that agrees with it. Officially:

- (iv) Definition: Suppose we are given a sentential language over a given set of variables. Let $Prob$ map formulas $\varphi_1 \cdots \varphi_k$, and pairs of formulas $(\chi_1, \psi_1) \cdots (\chi_j, \psi_j)$, into numbers. Then $Prob$ is coherent just in case there is a distribution \Pr (over the states arising from the variables) such that for all $i \leq k$,

$$Prob(\varphi_i) = \Pr(\varphi_i),$$

and for all $i \leq j$,

$$Prob(\chi_i \mid \psi_i) = \Pr(\chi_i \mid \psi_i).$$

If there is no such distribution, then $Prob$ is incoherent.

For example, the judgments in (iii) are coherent. This is because they agree with the distribution \Pr shown in (ii). In contrast, the following modification of the judgments is incoherent.

$$\begin{aligned} Prob'(p) &= .5 \\ Prob'(q \wedge \neg r) &= .25 \\ Prob'(\neg p \vee q) &= .20 \\ Prob'(q \wedge \neg r \mid p) &= .3. \end{aligned}$$

It should be clear that every state that makes true $q \wedge \neg r$ also makes true $\neg p \vee q$ (since $q \wedge \neg r$ implies $\neg p \vee q$). It is therefore impossible for a distribution to assign lower probability to the latter than to the former. Hence, no distribution agrees with $Prob'$.

The optimization problem

Assume that we are given a set of judgments represented by a mapping $Prob$ from formulas, and pairs of formulas, to numbers. If $Prob$ is incoherent, we seek to replace its values with coherent probabilities. Moreover, we seek replacements that best approximate the original values, so as to minimally distort the judge's opinions. The rationale behind the policy of minimal distortion is respect for the judge. Albeit incoherent, her assessments of chance might harbor insight into the uncertainty present in the environment. Minimally changing her numbers is the most plausible route to coherent judgment that still embodies her knowledge.

We measure the distance between two assessments of chance by their absolute difference, since this is the simplest and most interpretable measure. Other potential measures include the squared difference, or some version of relative entropy (which is not, however, a true distance; Cover and Thomas, 1991). None of the results reported below are substantially affected by use of these alternative measures. We thus have the following optimization problem.

- (v) Optimization Problem: Let $Prob$ map formulas $\varphi_1 \cdots \varphi_k$, and pairs of formulas $(\chi_1, \psi_1) \cdots (\chi_j, \psi_j)$, into numbers. Find a map $Prob^*$ with the same domain as $Prob$ such that $Prob^*$ is coherent, and

$$\sum_{i \leq k} | Prob(\varphi_i) - Prob^*(\varphi_i) | \quad + \quad \sum_{i \leq j} | Prob(\chi_i \mid \psi_i) - Prob^*(\chi_i \mid \psi_i) |$$

is minimized.

Note that this formulation assigns equal importance to approximating absolute and conditional probability judgments.

It is possible for there to be no minimum distance between $Prob$ and a coherent approximation $Prob^*$ to it. Consider, for example, the incoherent judgments $Prob(p \mid q) = .5$, $Prob(q) = 0$. (They are incoherent because the conditioning event q for $(p \mid q)$ has been assigned zero probability.) They can be approximated by setting $Prob^*(p \mid q) = .5$ and $Prob^*(q)$ arbitrarily close to 0, but not 0 itself so there is no best approximation. In all

cases, however, the minimum is bounded by zero. So we interpret (v) as requesting a coherent approximation $Prob^*$ to $Prob$ that is as close as possible within some positive tolerance.

Optimization via genetic algorithm

The main difficulty in solving the optimization problem (v) is combinatorial explosion. With n variables there are 2^n states, all of which may potentially interact with the coherence of a proposed approximation to $Prob$. The simplest means of handling large numbers of states is to limit the search to sparse distributions, i.e., to distributions in which many states have probability zero. Sparse distributions have compact representations since it is only necessary to encode the states with positive probability. They are thus easy to manipulate. This search strategy is the natural counterpart to the psychological hypothesis (i), which credits human reasoners with the ability to manipulate relatively few potential states-of-affairs at one time.⁴

We now describe a simple technique for finding a sparse distribution that approximates an input set of probability assessments. For concreteness, suppose the situation to be modeled involves 3 variables, p, q, r . Let M be any $3 \times m$ matrix all of whose entries are drawn from $\{t, f\}$ (truth and falsity). Then every column i of M represents a state, namely, the one that assigns $M(1, i)$ to p , $M(2, i)$ to q , and $M(3, i)$ to r . (The same state can be represented by more than one column.) Letting $m = 12$, one such matrix M is as follows.

(vi)

	1	2	3	4	5	6	7	8	9	10	11	12
p	t	f	t	f	t	f	t	f	t	f	t	f
q	f	f	t	f	t	f	f	f	t	f	f	f
r	t	t	t	f	t	f	t	t	t	t	t	t

In this example, column (1) of M represents the state in which p and r are true, and q is false. Column (4) represents the state in which all three variables are false. For each of the eight possible states s , we take M to assign s the probability n/m , where n is the number of columns of M that code s (and m is the total number of columns). Thus, in example (vi), the state in which all three variables are false has probability $2/12$ since just columns (4) and (6) represent it. Similarly, the state in which p, r are true and q false is assigned probability $3/12$. The state in which p is false and q, r are true has zero probability since it is not represented in the matrix. This way of interpreting M defines a (coherent) distribution over p, q, r . The same idea applies to any number of variables and columns.

Why does the distribution represented by M tend to be sparse? The number of states increases exponentially in the number n of variables. When the number of columns in M is modest, there will necessarily be states that do not correspond to a column and are thus assigned zero probability. Conversely, the greater the number of columns in M , the more states can receive nonzero probability, and hence the more distributions can potentially be represented within it.

In searching for a distribution that best approximates a corpus of judged probabilities, we limit attention to distributions that can be represented by matrices like (vi), using a fixed number of columns whose value is chosen to keep the problem feasible. Even for a modest number of columns, a large search space remains. (For just 3 variables and 12 columns, as in (vi), there are more than 68 billion possible matrices.) We propose to explore the space by interpreting a matrix M as a two-dimensional binary genome within the context of genetic algorithms (Michalewicz, 1994; Mitchell, 1996). Such algorithms include a population of chromosomes, each of which represents a potential solution to the search problem. They also rely on a scheme for evaluating the “fitness” of chromosomes in terms of their value as potential solutions. Fitness determines the probability that a given chromosome will participate in reproduction and thus help constitute the successor population of chromosomes (the “next generation”). The reproductive act includes crossover between the two chromosomes at a randomly determined point, as well as random mutation.

These concepts take the following form in our search context. With respect to a target corpus of judgments, M 's genetic fitness is measured in terms of the summed, absolute deviation between the probabilities that M assigns to the judged events (or pairs of events) and the original estimates of the judge, as stated in (v). A genetic algorithm designed to breed the matrix of greatest fitness will thus seek to construct a matrix-defined-distribution that best approximates the original judgments. Cross-over between two matrices exchanges sequences of columns (rather than rows). This allows each chromosome to be conceived as a string of states. Mutation flips a given cell of the matrix from one truth-value to the other. More details will be provided after we discuss the data on which the genetic algorithm was designed to operate.

Experimental test of the method

To determine the accuracy and computational feasibility of our approximation scheme, we elicited probability estimates from undergraduates and then brought them into coherence via a genetic algorithm.

Materials

Since weather and climate are common topics of conversation, they provide a domain in which many people can reveal insight (if not coherence). Our questions about this domain were built from the following sentential variables, 10 in all. Each sentence represents a weather forecast for noon, one week from the day on which the question is answered.

- (a) It is overcast in New York.
- (b) It is at least 56 degrees in New York.
- (c) It is overcast in Philadelphia.
- (d) It is at least 58 degrees in Philadelphia.
- (vii) (e) It is overcast in Houston.
- (f) It is at least 80 degrees in Houston.
- (g) It is overcast in Galveston.
- (h) It is at least 78 degrees in Galveston.
- (i) It is overcast in Los Angeles.
- (k) It is at least 68 degrees in Los Angeles.

Galveston is a Gulf Coast city 40 miles south of Houston, known to all participants in the study. Based on these ten variables, there are ninety complex events of each of the following six types, excluding cases in which the same variable is repeated.

- (a) Conditional statements of form p assuming-that q , such as “It is at least 78 degrees in Galveston assuming that it is overcast in Houston.”
- (b) Conditional statements of form p assuming-that $\neg q$, such as “It is at least 58 degrees in Philadelphia assuming that it is not overcast in New York.”
- (c) Conjunctions of form $p \wedge q$, such as “It is overcast in Galveston and it is overcast in Houston.”
- (viii) (d) Conjunctions of form $p \wedge \neg q$, such as “It is overcast in New York and it is less than 58 degrees in Philadelphia.”
- (e) Disjunctions of form $p \vee q$, such as “It is at least 78 degrees in Galveston or it is overcast in Los Angeles.”
- (f) Disjunctions of form $p \vee \neg q$, such as “It is overcast in Houston or it is less than 80 degrees in Houston.”

Note that the negation of “at least x degrees” is expressed as “less than x degrees.” All the probability estimates requested in the study were drawn from the 10 elementary events in (vii), and the $6 \times 90 = 540$ complex events described in (viii).

Participants

Thirty-eight Rice University undergraduates participated in the study. They were unpaid volunteers, fulfilling a course requirement.

Procedure

It was explained to each student that they were to provide probability estimates for various meteorological events. The time of occurrence of all events was noon, one week from the day of the experiment, local time (e.g., noon in Los Angeles). All the events would involve the ten statements shown in (vii), which were presented. A map of the United States was also presented, with the five cities indicated.

It was pointed out that “overcast” meant either cloudy, partly cloudy, rainy, or snowy, but not hazy. Examples of complex events were then presented, and the intuitive meaning of conditional probability (embodied by the expression “assuming that”) was reviewed. It was further noted that “or” was used in the inclusive sense; for example, (viii)e would be true if either it is at least 78 degrees in Galveston or it is overcast in Los Angeles, or both. Students were then directed to a website on which they would enter their probability estimates sometime during the day. It was stated that the accuracy of their forecasts would be computed (using a “standard measure”) and the most accurate forecaster would be given a prize. Forecasts would be verified by recourse to the CNN weather webpage on the day in question. Finally, subjects were advised that there would be 46 estimates, and to pace themselves to avoid fatigue.

The website reminded the student of the points raised at the earlier meeting, then presented 46 events for probability estimation. The first ten were an individually randomized ordering of the elementary events (vii). For each of the six classes of complex events shown in (viii), six events (or pairs of events in the conditional cases) were randomly chosen individually for each participant. The resulting 36 complex events were then presented in individually randomized order with the restriction that the six events in each class be presented as a block. All responses were constrained by the electronic questionnaire to fall in the interval $[0, 1]$.

The study was completed during October and the first part of November, 1999. From one to five students forecast weather for the same day (typically, the number was two or three).

Results

Incoherence of judgment

The third column of Table 1 shows the mean and standard deviation of the average probability estimates for each of the seven types of events, averaging over all 38 participants. In conformity with the probability calculus, the estimates made for elementary events tend to be greater than those for conjunctions and less than those for disjunctions. Nonetheless, the students’ judgments showed marked incoherence, as we will now see.

Suppose that *Prob* represents the assessments of a given subject. It is easy to show the following fact.

(ix) Let p and q be sentential variables. Then necessary conditions on the coherence of $Prob$ include:

- (a) $Prob(p) + Prob(q) - 1 \leq Prob(p \wedge q) \leq \min\{Prob(p), Prob(q)\}$
- (b) $Prob(p) - Prob(q) \leq Prob(p \wedge \neg q) \leq \min\{Prob(p), 1 - Prob(q)\}$
- (c) $\max\{Prob(p), Prob(q)\} \leq Prob(p \vee q) \leq Prob(p) + Prob(q)$
- (d) $\max\{Prob(p), 1 - Prob(q)\} \leq Prob(p \vee \neg q) \leq 1 + Prob(p) - Prob(q)$
- (e) $Prob(p | q) = Prob(p \wedge q) / Prob(q)$
- (f) $Prob(p | \neg q) = Prob(p \wedge \neg q) / Prob(\neg q)$.

Every student had six occasions to violate each of (ix)a-d. Of these six occasions, the mean number of violations of (ix)a-d was 2.97, 3.00, 2.08, and 2.95, respectively.

There were fewer occasions to violate (ix)e,f because conjunctions and conditional probability items were sampled independently for each participant. Thus, it was possible for a given student to supply $Prob(p | q)$ but not $Prob(p \wedge q)$. The same is true for $Prob(p | \neg q)$ and $Prob(p \wedge \neg q)$. Across all 38 students, there were 55 occasions to violate either (ix)e or (ix)f. Such violation occurred 51 times. These results are summarized in the first three columns of Table 2.

Incoherence may be due in part to numerical imprecision in judgment. This is especially true for violations of conditions (ix)e,f, which are equalities. Incoherencies were therefore retabulated, this time loosening the constraints by a tolerance of .01. For example, in place of (ix)a, an estimate was considered coherent if it satisfies:

$$Prob(p) + Prob(q) - 1.01 \leq Prob(p \wedge q) \leq \min\{Prob(p), Prob(q)\} + .01.$$

Similarly, to be counted as coherent, conditional probabilities had to be within .01 of their defining quotient. Such tolerance had little effect on assessed incoherence, as shown in the fourth column of Table 2. The last column of the table reveals that increasing the tolerance to .05 also has small effect.

Accuracy of the students' estimates

Quadratic score

If a given meteorological event comes true, it is natural to assign it probability 1, and to assign it 0 otherwise. A qualification is needed in the case of conditional events like "overcast in Houston assuming at least 56 degrees in New York." If the conditioning event is false (i.e., it is less than 56 degrees in New York) then no value should be assigned to the conditional. Probabilistic forecasts can then be compared to these numbers in order to determine forecast accuracy. A common metric of comparison is the "quadratic score" (von Winterfeldt and Edwards, 1986), defined as follows.

- (x) Definition: Suppose that *Prob* represents the assessments of a given judge. Let E be an event in the domain of *Prob*, and let (G, F) be a pair of events in the domain of *Prob*.
- (a) The quadratic score incurred by *Prob* for E is $(1 - \text{Prob}(E))^2$ if E is true. It is $\text{Prob}(E)^2$ if E is false.
- (b) The quadratic score incurred by *Prob* for the pair (G, F) is $(1 - \text{Prob}(G \mid F))^2$ if both G and F are true. It is $\text{Prob}(G \mid F)^2$ if G is false and F is true. It is not defined if F is false.

The quadratic score of *Prob* is the average of all the penalties incurred by *Prob* for events and pairs of events in its domain. (Pairs of events for which the quadratic score is not defined do not figure in this average.)

To illustrate, suppose a judge assigns .3 probability to the disjunction “either overcast in Philadelphia or overcast in New York.” If it is overcast in either of the cities, the disjunction is true so the score is $(1 - .3)^2$. If it is overcast in neither city, the disjunction is false so the score is $.3^2$. Suppose $\text{Prob}(\text{It is overcast in Galveston} \mid \text{It is overcast in Houston}) = .4$, and it turns out not to be overcast in Houston. Then the score associated with this pair of events is not defined. Note that the score is a penalty. Hence, lower scores reveal more insight than higher ones.

The quadratic score is a popular measure of judgmental accuracy for the following reasons. When judges make a conscious effort to minimize their score, the quadratic rule encourages honest assessments of chance (unlike the use of absolute difference, which yields lower expected penalties if estimates are sharpened towards 0 and 1).⁵ It also decomposes in a revealing way into coefficients that can be extracted from many assessment contexts (Murphy, 1973; for discussion, see Yates, 1990).

A judge who feels entirely ignorant about a specific assessment will likely respond with .5. Systematic use of this strategy leads to incoherence, since not all of $p \wedge q$, $p \wedge \neg q$, $\neg p \wedge q$, $\neg p \wedge \neg q$ can have probability .5. We may nonetheless conceive of a totally ignorant judge faced with just one assessment, and average the resulting penalties over all such estimates. The average quadratic score of a judge who responds .5 is .25, so .25 serves as an appealing threshold of ignorance for this score. A judge with lower score gives evidence of insight into the events under scrutiny; higher scores indicate lack of insight, or worse (scores close to 1 reflect inverted insight).

For each of the 38 participants, we determined the relevant meteorological events for the day whose weather was to be probabilistically forecast. We then calculated individual quadratic scores. Of the 46 estimates provided by each student, 12 were conditional probabilities. The conditioning events were not always true, so the scores of a given participant was not always based on 46 estimates. In fact, the average number of estimates on which the quadratic score was computed is 39.6 (s.d. = 1.58).

The mean quadratic score for the 38 students is .231 (s.d. = .056). By t -test, this value is reliably less than the .25 ignorance threshold ($p < .05$, two-tailed). Twenty-four of the 38 participants had a score below .25. Mean penalties are broken down by type of judgment in the third column of Table 3.

On the average, the students showed insight about the weather. The insight is limited, but nonetheless impressive because it bears on meteorological events one week hence (too long to favor prediction). Moreover, their insight shines through the ample incoherence seen above, which limits accuracy. Despite the incoherence, the students' judgments provide a guide to the weather that is reliably better than retreating to the estimate of .5 for a given event.

Slope of the judgments

For each subject we isolated the cases in which the judged event came true (or, in the conditional case, both the conditioning event and the target event came true). The average probability assigned to these cases was compared to the average probability assigned to events that did not come true (or, in the conditional case, assigned to target events that did not come true even though the conditioning event did). The difference between these means is called the slope of the judgments, and measures the judge's ability to distinguish the truth and falsity of predictions (high slope reflects prescience; see Yates, 1990, Ch. 3).

The mean slope over the 38 subjects was .154 (s.d. = .135), reliably greater than zero ($t(37) = 7.03$, $p \leq .001$). This mean will shortly be compared to the slopes obtained from the coherent approximations of the students' judgments.

Coherent approximations

Two indices will be used to measure how well a given distribution approximates a set of probability estimates. One measure is the mean absolute deviation (MAD) between the judge's estimate of the chance of a given event and the distribution's value for the same event. The second measure is the Pearson correlation between these numbers, once again taken over all 46 assessed events. A good approximation to a judge is a distribution that generates a low MAD and a high correlation. Incoherent estimates cannot be perfectly approximated by coherent substitutes. The considerable incoherence shown by our students thus imposes a lower bound on the MAD that can be achieved by an approximating distribution.

Uniform distribution

The crudest means of approximating a set of incoherent judgments is to take no account of the particularities of the judge, and simply replace her assessments with the probabilities of some fixed distribution. As a baseline measure of performance, we carried out this procedure with the uniform distribution, which assigns the same probability (namely, $1/1024$)

to every state generated by our 10 variables. According to the uniform distribution, the probability of each elementary event is $1/2$, conditional probabilities are $1/2$, conjunctions have probabilities $1/4$, disjunctions $3/4$. Over the 38 participants, the average MAD produced by uniform approximation is .183 (s.d. = .044). The average correlation between the subjects' judgments and those issuing from the uniform approximation is .434 (s.d. = .221). This level of correlation is reliably greater than zero ($N = 46$, $p \leq .01$). The low level of correlation suggests that the judges were appropriately sensitive to more than the logical structure of the events they were evaluating (since the uniform distribution assigns probabilities solely on the basis of logical structure).

Distributions based on independence

The uniform distribution is one of a family of distributions that enforce independence among the variables. That is, it conforms to:

- (xi) Independence Property: For variables p, q, r , $\Pr(p \wedge q \wedge r) = \Pr(p) \times \Pr(q) \times \Pr(r)$, and similarly for other conjunctions.

If a distribution satisfies (xi), it is determined by specifying the probabilities of the elementary events; all remaining probabilities and conditional probabilities may be derived on this basis. For each participant there is thus a unique distribution satisfying (xi) and agreeing with her estimates of elementary events. Use of this distribution to approximate a set of estimates has the merit of sensitivity to at least some of the judge's opinions (namely, for elementary events). Since it pays no attention to the remaining assessments, the distribution is the crudest approximation that is individually tailored to the judge. It thus serves as a second baseline for the accuracy of coherent approximations.

For each of the 38 participants we calculated the MAD and the correlation associated with the distribution that satisfies (xi) and agrees with the probabilities assigned to elementary events. The average MAD obtained was .144 (s.d. = .049). The average correlation was .625 (s.d. = .193). A correlated t -test reveals the new approximations to be reliably better than those based on the uniform distribution ($t(37) = 6.9, 8.2$, respectively, $p \leq .001$). That judgments were not perfectly correlated with the independent distribution shows that the judges were appropriately sensitive to the interdependence of meteorological events.

Distributions constructed by genetic algorithm

For each student we constructed a population of 200 chromosomes. Each chromosome consisted of 10 rows (representing the 10 variables) and 100 columns (representing some of the 1024 states); see the earlier discussion. The 1000 cells of a given chromosome were randomly filled with t and f (truth and falsity) under the constraint that the percentage of

t 's in row i equal the probability that the student assigned to variable i (except for rounding error). Thus, every chromosome in the starting population reflected the probabilities the student assigned to the elementary events. The 200 chromosomes were then evolved through 500 generations. Fitness of a given chromosome was measured by $1/(\text{MAD} + .01)$, where MAD is the mean absolute deviation between the student's estimates and those embodied by the chromosome (adding .01 prevents division by zero). Between two generations, the probability of being selected for mating was proportional to fitness. One hundred such pairs were selected (with replacement) on this basis. With probability .5 the pair underwent "crossover" at a randomly chosen column. Whether crossed or not, the pair then underwent "mutation" in the form of a .001 probability of flipping any given cell. The pair of chromosomes then entered the next generation (again yielding a population of 200). In this procedure, fitness was calculated 100,000 times (i.e., 200 times per generation). The chromosome with highest fitness (lowest MAD) was retained, and its distribution was used to approximate the student's 46 assessments of chance. The entire procedure is performed for one subject at a time (there is no interaction between chromosomes generated for different students).

The average MAD achieved on this basis was .114 ($N = 38$, s.d. = .043). A correlated t -test reveals this performance to be reliably better than that achieved with distributions based on independence ($t(37) = 13.45$, $p \leq .001$). (Hence, it is also better than use of the uniform distribution.) Indeed, for all 38 students, the MAD for the genetic algorithm was less than that using independence. The average correlation achieved via the genetic algorithm is .711 ($N = 38$, s.d. = .185; each correlation involves 46 pairs). The correlations of the genetic algorithm are reliably higher than for distributions based on independence ($t(37) = 8.67$, $p \leq .001$). This improvement was seen for 37 of the 38 students.

The fourth column of Table 1 shows the mean probabilities assigned by the 38 coherent approximations to judgments of different types.

Accuracy of the reconstructed estimates

For each participant we calculated the quadratic score for the 46 estimates derived from her best-approximating chromosome. Overall, the mean score for the reconstructed estimates is .204 (s.d. = .047) compared to .231 (s.d. = .056) for the original estimates, as reported earlier. By a correlated t -test, the difference is reliably different from zero ($t(37) = 6.92$, $p \leq .001$). (The difference from the ignorance threshold of .25 is also reliable.) For 33 of the 38 participants, the reconstructed estimates had a lower quadratic score than did the original estimates. The coherent revisions improve the accuracy of forecasts for each of the seven types of events in the study. Table 3 compares the average mean quadratic score for the students versus their coherent approximations with respect to each type of judgment. Improvements are seen in every case, reaching statistical significance about half the time (by correlated t -test). It should be borne in mind that improving accuracy via coherent approximation to judgment is not a mathematical necessity. For any objective

state of affairs, many coherent approximations to a given set of judgments make their quadratic score worse.

The coherent approximations also improved the discrimination of true from false statements. Across the 38 subjects, the average slope of the coherent approximations was .203 (s.d. = .106). By correlated t -test this is reliably higher than the .154 slope achieved by the original judgments ($t(37) = 5.82, p \leq .001$).

The accuracy improvements reported above are reliable but modest. For example, the average reduction of quadratic score corresponds to assigning a probability of .45 rather than .48 to a false proposition. The point of the analysis, however, is only to show that our coherent approximations do not diminish the accuracy of the judge. This seems to be the case inasmuch as both accuracy and slope reliably increased.

Discussion

A judge whose estimates are incoherent is likely to demand two things of a coherent revision. First, the distortion of her original estimates should be minimal. Second, no loss of predictive accuracy should result from the revision.

Regarding the first desideratum, use of the genetic algorithm yields coherent approximations to the students' judgments that were reliably closer than either of our baseline techniques. The latter were (a) use of the uniform distribution over states for every participant, and (b) construction of an individually tailored distribution starting from the student's probabilities for elementary events and assuming independence. Using our algorithm, the average MAD achieved was .114, and the average correlation between original and revised assessments was .711. These results were achieved despite considerable incoherence in the participants' assessments, which limits the closeness of any coherent approximation. The relative success of our method supports the hypothesis of sparse distributions (i) according to which good coherent approximations to human judgment can be found among distributions assigning positive probability to few potential states-of-affairs. This is because the genetic algorithm delivers only sparse distributions of probability as (coherent) approximations to judgment.

It is noteworthy that computing the 500 generations produced by our genetic algorithm took only a few minutes per student on a personal computer. Most of the progress occurred within the first two hundred generations, further descent occurring more and more intermittently thereafter. (When only a few generations or a small population of chromosomes are employed, the genetic algorithm delivers a coherent approximation that is no closer to the original judgments than the distribution based on independence.) The ease-of-computation seen in the present study bodes well for scaling our technique to larger problems, involving more than 10 variables.

Judges might be more confident about certain estimates compared to others, and desire greater fidelity to them in a coherent revision. For example, conditional probabilities are often more psychologically accessible than absolute ones. (It seems easier to estimate the

chance of rain in Minneapolis given rain in St. Paul than to estimate the chance of either event alone.) In the genetic algorithm, fitness can be defined to take account of preferences among estimates. For example, the absolute distance between input probabilities and those coded by a chromosome may be weighted by a coefficient reflecting the judge's confidence in the estimate. For simplicity in the present study, no such weighting was imposed.

The chromosomes of our genetic algorithm were evaluated for fitness by comparing them to numerical estimates of probability and conditional probability. A variety of alternative judgments can also be used to evaluate fitness. For example, judges might specify (a) inequalities among the chances of events, (b) conditional independence among variables, (c) correlation among events. Provided that a given distribution can be evaluated for its proximity to an input judgment, the method of genetic algorithms allows the judgment to guide the search for a coherent approximation.

The judge's second desideratum is that her revised estimates be no less accurate than the original ones. Accuracy was measured via the quadratic score. Not only was accuracy undiminished, the coherent approximations had reliably better scores compared to the original judgments (see Table 3). This phenomenon is similar to "bootstrapping" in the prediction of quantitative variables like college grades. A linear model of the judge's estimates is often a better predictor than the judge herself.⁶ Bootstrapping probabilities has an aspect not found in the context of linear models, however. Whereas there are no normative grounds for using a linear model to predict college grades from SATs, high school grades, etc., there are persuasive reasons to prefer one's probabilities to be coherent (see the earlier discussion about utility analyses and susceptibility to "Dutch books").

Probabilistic bootstrapping is most useful when the events in question cannot be easily assimilated to a large class of similar instances. Geopolitical forecasting is an example of such a situation since the probability, e.g., of Switzerland entering the European Union before 2020 cannot be extrapolated from a class of similar historical moments. Human judgment must be relied upon to estimate the chances of these kinds of events, which opens the door to incoherence. In the contrary case, when probabilities can be extrapolated from past data, the resulting set of estimates is guaranteed to be coherent (provided that relative frequencies are calculated from the same dataset for all the events in play).

Finally, we note that the present technique offers a method of aggregating the opinions of a panel of experts asked to assess the chances of the same events. (For an overview of issues and methods for aggregating judgment, see Ferrell, 1985; Rowe, 1992.) This problem is typically studied in the context of elementary events, like (vii) (see, for example, Ariely et al., 2000). In a more general context, experts may be asked to assess overlapping sets of complex and elementary events. Even if the assessments of a given expert are coherent, the union (i.e., combined set) of two experts' assessments is unlikely to be so. To extract a single set of coherent estimates from the panel, one approach is to take the union of all the judgments and find the best coherent approximation using a genetic algorithm (or some other approximation method). The result will be a compromise distribution that takes

everyone's views into account, and distorts them minimally. If some members of the panel have better credentials than others, their estimates can be weighted more heavily in the process.

References

- Alpert, M. and Raiffa, H. (1982). A progress report on the training of probability assessors. In Kahneman, D., Slovic, P., and Tversky, A., editors, *Judgment under uncertainty: Heuristics and biases*, pages 294–305. Cambridge University Press, New York NY.
- Ariely, D., Au, W. T., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., Wallsten, T. S., and Zauberman, G. (2000). The Effects of Averaging Subjective Probability Estimates Between and Within Judges. *Journal of Experimental Psychology: Applied*, 6(2):130–147.
- Bahar, R., Frohm, E., Gaona, C., Hachtel, G., Macii, E., Pardo, A., and Somenzi, F. (1997). Algebraic decision diagrams and their applications. *Journal of Formal Methods in Systems Design*, 10(2/3):171–206.
- Baron, J. (1994). *Thinking and Deciding (2nd edition)*. Cambridge University Press, New York NY.
- Bernardo, J. and Smith, A. (1994). *Bayesian Theory*. John Wiley and Sons, New York.
- Bradley, R. (1999). More triviality. *Journal of Philosophical Logic*, 28(2):129–139.
- Camerer, C. (1981). General conditions for the success of bootstrapping models. *Organizational Behavior and Human Performance*, 27:411–422.
- Castillo, E., Gutiérrez, J., and Hadi, A. (1997). *Expert systems and probabilistic network models*. Springer, New York.
- Coletti, G. (1990). Coherent Qualitative Probability. *Journal of Mathematical Psychology*, 34:297–311.
- Coletti, G., Gilio, A., and Scozzafava, R. (1993). Comparative probability for conditional events: A new look through coherence. *Theory and Decision*, 35:237–258.
- Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. John Wiley & Sons, New York NY.
- Dawes, R. M. (1979). The robust beauty of improper linear models. *American Psychologist*, 34:571–582.
- Dawes, R. M. and Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, 81:97–106.

de Finetti, B. (1972). *Probability, Induction, and Statistics*. Wiley, New York.

Druzdzel, M. J. and van der Gaag, L. C. (1995). Elicitation of probabilities for belief networks: Combining qualitative and quantitative information. In *Uncertainty in Artificial Intelligence (95): Proceedings of the 11th conference*, Los Altos CA. Morgan Kaufmann.

Ferrell, W. (1994). Discrete subjective probabilities and decision analysis: elicitation, calibration and combination. In Wright, G. and P. Ayton, P., editors, *Subjective Probability*. John Wiley and Sons, New York NY.

Georgakopoulos, G., Kavvadias, D., and Papadimitriou, C. (1988). Probabilistic satisfiability. *Journal of Complexity*, 4:1–11.

Gustason, W. (1994). *Reasoning from Evidence: Inductive Logic*. Macmillan, New York NY.

Henrion, M. (1987). Practical issues in constructing a Bayes' belief network. *Third Workshop in Uncertainty and Artificial Intelligence*, pages 132–139.

Holtzman, S. and Breese, J. (1986). Exact reasoning about uncertainty: On the design of expert systems for decision support. In Lemmer, J. F. and Kanal, L. N., editors, *Uncertainty and Artificial Intelligence*, pages 339–346. North-Holland/Elsevier, New York NY.

Jeffrey, R. C. (1983). *The Logic of Decision (2nd Edition)*. The University of Chicago Press, Chicago IL.

Klayman, J. and Brown, K. (1993). Debias the environment instead of the judge: an alternative approach to reducing error in diagnostic (and other) judgment. *Cognition*, 49:97–122.

Kleindorfer, P. R., Kunreuther, H. C., and Schoemaker, P. J. (1993). *Decision Sciences: An Integrative Perspective*. Cambridge University Press, Cambridge, England.

Kleinmuntz, D. N., Fennema, M. G., and Pecher, M. E. (1996). Conditioned Assessment of Subjective Probabilities: Identifying the Benefits of Decomposition. *Organizational Behavior and Human Decision Processes*, 66(1):1–15.

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *Philosophical Review*, 85:297 – 315.

Lindley, D. V., Tversky, A., and Brown, R. V. (1979). On the reconciliation of probability assessments. *Journal of the Royal Statistical Society A*, 142 (Part 2):146–180.

Manktelow, K. (1999). *Reasoning and Thinking*. Psychology Press, East Sussex, UK.

Michalewicz, Z. (1992). *Genetic Algorithms + Data Structures = Evolution Programs*. Springer-Verlag, Berlin.

Mitchell, M. (1996). *An introduction to genetic algorithms*. MIT Press, Cambridge MA.

Morgan, M. G. and Henrion, M. (1990). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, Cambridge, England.

Murphy, A. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12:595–600.

Neapolitan, R. (1990). *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. John Wiley & Sons, New York NY.

Nilsson, N. (1986). Probabilistic logic. *Artificial Intelligence*, 28(1):71–87.

Osherson, D. (1995). Probability judgment. In Smith, E. E. and Osherson, D., editors, *Invitation to Cognitive Science: Thinking (Second Edition)*. M.I.T. Press, Cambridge MA.

Osherson, D., Shafir, E., Krantz, D., and Smith, E. E. (1997). Probability bootstrapping: Improving prediction by fitting extensional models to knowledgeable but incoherent probability judgments. *Organizational Behavior and Human Decision Processes*, 69(1):1–8.

Osherson, D., Shafir, E., and Smith, E. E. (1994). Extracting the coherent core of human probability judgment. *Cognition*, 50:299 – 313.

Osherson, D., Smith, E. E., Shafir, E., Gualtierotti, A., and Biolsi, K. (1995). A Source of Bayesian Priors. *Cognitive Science*, 19(3):377 – 405.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.

Resnik, M. D. (1987). *Choice: An introduction to decision theory*. University of Minnesota Press, Minneapolis MN.

Rowe, G. (1992). Perspectives on expertise in aggregation of judgments. In Wright, G. and F. Bolger, P., editors, *Expertise and Decision Support*, pages 155–180. Plenum Press, New York NY.

von Winterfeldt, D. and Edwards, W. (1986). *Decision analysis and behavioral research*. Cambridge University Press, New York NY.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, England.

Walley, P. (1996). Measures of uncertainty in expert systems. *Artificial Intelligence*, 83:1–58.

Yates, J. F. (1990). *Judgment and Decision Making*. Prentice Hall, Englewood Cliffs NJ.

Notes

(1) The computational complexity of probabilistic coherence is discussed in Georgakopoulos, Kavvadias and Papadimitriou (1988). The human tendency to stray into incoherent estimates of chance is reviewed in Yates (1990), Baron (1994), and Osherson (1995). It is striking to observe, for example, how few people realize that it is inconsistent to attribute probabilities of .8 to each of two sentences, and probability .5 to their conjunction; see (ix)a, below.

(2) For discussion of coherence and Dutch Books, see Osherson (1985), Resnik (1987), Coletti (1990), Coletti, Gilio and Scozzafava (1993), Gustason (1994). In fact, if judgments are incoherent, then bets can be chosen so that each has positive expected value according to the judge yet are collectively guaranteed to lead to a loss.

(3) See Alpert and Raiffa (1982), Holtzman and Breese (1986), von Winterfeldt and Edwards (1986), Henrion (1987), Morgan and Henrion (1990), Klayman and Brown (1993), Druzdzel and van der Gaag (1995). The latter paper, for example, describes how elicitation techniques can be extended to nonnumerical judgments of independence and conditional independence. A separate issue from ensuring coherence is improving the accuracy of probability estimates. Elicitation techniques that decompose simple probabilities via the law of total probability have been shown to sometimes lead to more accurate estimates of chance. See Kleinmuntz, Fennema and Peecher (1996) for experimental results and discussion.

(4) We note that sparse distributions are not the only means of compactly encoding coherent probabilities. Some non-sparse distributions, for example, can be factored into smaller subdistributions that interact via multiplication; see Pearl (1988), Neapolitan (1990). This approach relies on assumptions about conditional independence, however, that may not be realistic in practice. Yet other classes of distributions can be compactly described via algebraic decision diagrams, in the sense of Bahar et al. (1997). It is also worth noting that the use of a quadratic objective function in place of (v) does not render the optimization problem any easier. For example, it can be shown that the quadratic counterpart to (v) is nonconvex on its domain.

(5) That is, if a judge believes that the probability of an event is p , and announces the probability as q , then she minimizes her expected quadratic score—namely $[p \times (1 - q)^2] + [(1 - p) \times q^2]$ —by setting $q = p$ (see Bernardo and Smith, 1994, Section 2.7). In contrast, setting $q = p$ does not minimize the expected absolute score, namely, $[p \times (1 - q)] + [(1 - p) \times q]$. Of course, our subjects were not informed about any scoring procedure, and had no reason whatsoever to falsify their probability estimates.

(6) See Dawes (1979), Dawes and Corrigan (1974), Camerer (1981). Extension of the bootstrapping concept to probabilistic estimates is discussed in Osherson, Shafir, Krantz and Smith (1997). See also Osherson, Shafir and Smith (1994).

Author notes

We thank David H. Krantz, Jonathon Baron and an anonymous reviewer for close reading of an earlier version of this paper, and helpful comments.

Author addresses: Rice University, P.O. Box 1892, Houston TX 77251-1892. E-mail: osherson/lane/batsell/hartley@rice.edu.

TABLE 1:
 Mean and Standard Deviation for the Seven Types of Estimates.

<i>Type</i>	<i>N</i>	Subject	Coh. approx.
		<i>Mean (s.d.)</i>	<i>Mean (s.d.)</i>
p	380	.547 (.088)	.529 (.064)
$p \mid q$	228	.451 (.127)	.488 (.090)
$p \mid \neg q$	228	.473 (.152)	.500 (.097)
$p \wedge q$	228	.411 (.140)	.311 (.104)
$p \wedge \neg q$	228	.367 (.143)	.263 (.038)
$p \vee q$	228	.687 (.157)	.769 (.069)
$p \vee \neg q$	228	.639 (.139)	.735 (.041)
overall	1748	.514 (.071)	.515 (.052)

Table legend: The first column shows the type of event whose probability was estimated. Elementary events are first, followed by the types listed in (viii). The second column shows the total number of judgments of each type evaluated by the subjects. The third column gives the mean for the subjects' average judgment. Thirty-eight subjects figure in this mean. The fourth column provides the same information for the subjects' approximating coherent distributions.

TABLE 2:
Average Number of Incoherent Weather Forecasts by the 38 Judges.

	N	0 <i>tol</i>	.01 <i>tol</i>	.05 <i>tol</i>
$p \wedge q$	6	2.97	2.89	2.68
$p \wedge \neg q$	6	3.00	2.55	2.42
$p \vee q$	6	2.08	2.08	1.87
$p \vee \neg q$	6	2.95	2.55	2.34
$p \pm q$	1.44	1.34	1.32	1.24

Table legend: The first column specifies a constraint on coherence, in the sense of (ix). The expression $p | \pm q$ signifies the combination of constraints (ix)e,f, namely, $Prob(p | q) = Prob(p \wedge q) / Prob(q)$ and $Prob(p | \neg q) = Prob(p \wedge \neg q) / Prob(\neg q)$. The second column shows the average number of occasions on which the test can be made. The average number of incoherent judgments is given in the third column. The fourth column provides the same information when incoherence of .01 is tolerated. A tolerance of .05 yields the data in the last column.

TABLE 3:
 Quadratic scores for students and for their coherent approximations.

<i>Type</i>	# of students	<i>Student Mean (s.d.)</i>	<i>Coh. approx. Mean (s.d.)</i>	signif.
p	38	.247 (.075)	.240 (.051)	.231
$p \mid q$	36	.270 (.156)	.265 (.114)	.748
$p \mid \neg q$	36	.257 (.158)	.234 (.122)	.157
$p \wedge q$	38	.243 (.107)	.192 (.090)	.001
$p \wedge \neg q$	38	.201 (.092)	.156 (.095)	.001
$p \vee q$	38	.207 (.141)	.187 (.139)	.278
$p \vee \neg q$	38	.212 (.105)	.178 (.097)	.053
all	38	.231 (.056)	.204 (.047)	.001

Table legend: The first column shows the type of event being judged. The second column shows the number of participants figuring in the mean. The number is less than 38 for conditional judgments since the conditioning event occasionally failed to be satisfied on any of the six occasions. The last column shows the two-tailed p -value associated with a correlated t -test of the hypothesis of equal means for students versus their approximations.