Aggregating Disparate Estimates of Chance*

Daniel Osherson
Princeton University

Moshe Y. Vardi
Rice University

August 15, 2004

Abstract

We consider a panel of experts asked to assign probabilities to events, both logically simple and complex. The events evaluated by different experts are based on overlapping sets of variables but may otherwise be distinct. The union of all the judgments will likely be probabilistic incoherent. We address the problem of revising the probability estimates of the panel so as to produce a coherent set that best represents the group's expertise.

Acknowledgements

We warmly thank three referees of an earlier version of this paper for insightful and detailed comments. The research reported herein was supported by NSF grant IIS-9978135.

*Contact: osherson@princeton.edu, vardi@rice.edu.

Introduction

Aggregating opinion

When deciding what to believe, it often makes sense to consult other people's opinions. To be useful, however, the opinions must be processed by some method for converting them into your ultimate judgment. Let us call such a method an *aggregation principle*. Formulating and justifying principles of aggregation pose unexpected challenges. Consider an example raised by Kornhauser & Sager (1986).[1] Three experts issue opinions about the four statements symbolized at the top of the following table.

(1)

Expert	p	q	$r \leftrightarrow (p \wedge q)$	r
1	agree	agree	agree	agree
2	agree	disagree	agree	disagree
3	disagree	agree	agree	disagree
Majority	agree	agree	agree	disagree

Each expert is logically consistent, and you would like to agree with a given statement if and only if the majority of experts do. But this turns out to be unadvisable, since the majority-opinion (shown at the bottom of the table) is logically inconsistent. This difficulty does not stem from the absence of one's own opinion when trying to integrate those of the experts. You can imagine that you are one of the experts, and even that your opinions are endowed with extra credibility. It is easy to see that a weighted voting scheme will nonetheless come to grief in certain cases.

A more promising response to Kornhauser & Sager's puzzle is to distinguish statements in terms of logical complexity. The experts can be polled just on p, q, r, leaving the status of $r \leftrightarrow (p \wedge q)$ to be deduced. Applied to the table, this policy yields agreement with p, q and disagreement with $r, r \leftrightarrow (p \wedge q)$, which is a consistent point of view. To generalize this approach, a few definitions are necessary.

By a *polarity variant* of a set $\{\varphi_1 \cdots \varphi_m\}$ of formulas we mean any set of form $\{\pm\varphi_1 \cdots \pm \varphi_m\}$, where $\pm\varphi_i$ is φ_i with one or zero occurrences of $\neg$ in front it. For example, $\{p, \neg(q \vee r)\}$ is a polarity variant of $\{p, (q \vee r)\}$. By a *basis* of a set A of formulas is meant a subset B of A that meets the following conditions.

(a) The members of B are *logically independent*, that is, every polarity variant of B is consistent.

(b) For every polarity variant B' of B and formula $\psi \in A$, either B' logically implies ψ or B' logically implies $\neg\psi$.

Intuitively, a basis of A is a minimal subset of A whose truth values suffice to pin down the truth values of the rest of A. For example, $\{p, q\}$ is a basis for $\{p, q, p \wedge q\}$. In contrast, the set $\{p \wedge q\}$ is not a basis for $\{p, q, p \wedge q\}$ because the polarity variant $\{\neg(p \wedge q)\}$ does not imply either p or $\neg p$ (and similarly for q). Thus, a promising strategy for aggregating multiple opinions about a given set of statements is to apply majority rule to some basis for the set, obtaining one of its polarity variants; the rest can be filled in via deduction. The resulting judgments will be consistent.

The foregoing scheme has defects, however. For one thing, not every consistent set of formulas has a basis. To illustrate, there is no basis for $\{p, p \wedge q\}$, as easily verified. Worse yet is the possibility of a surfeit of bases with no grounds for choosing among them. In particular:

(2) FACT: There is a consistent set A of formulas such that:

(a) A has at least three bases, and

(b) the bases of A have identical logical complexity.

Letting $A = \{p \leftrightarrow q, q \leftrightarrow r, p \leftrightarrow r\}$ witnesses Fact (2). The bases of A are:

(3) $\{p \leftrightarrow q, q \leftrightarrow r\}$ $\{p \leftrightarrow q, p \leftrightarrow r\}$ $\{q \leftrightarrow r, p \leftrightarrow r\}$

Indeed, every polarity variant of each of these sets determines the truth values of all three formulas in A. Plainly, the three bases have equivalent complexity along every dimension.

What does Fact (2) mean for aggregating multiple opinions? In a situation where there are a plurality of bases, a voting scheme can nonetheless be used provided that everyone agrees about which basis to use. Voting about the elements of the basis will then yield a convincing representation of group opinion. If there is disagreement about the basis, we might hope to resort to a syntactic criterion, notably, favoring the basis of least logical complexity (perhaps believing that judgment is better about simpler cases). But in view of Fact (2), syntax will not always decide the matter. The obvious strategy is then to ask each expert to rank order the bases in terms of her preference for their use in aggregation. But then *the problem of aggregating belief reduces to the problem of aggregating preference*, and hence is not open to entirely satisfactory solution. (For logical obstacles to aggregating preferences, see (Kelly 1978, Johnson 1998) among many other sources.)

We can illustrate with the set $\{p \leftrightarrow q, q \leftrightarrow r, p \leftrightarrow r\}$ again. Suppose that three experts announce the following (consistent) opinions.

(4)

Expert	$p \leftrightarrow q$	$q \leftrightarrow r$	$p \leftrightarrow r$
1	disagree	disagree	agree
2	agree	disagree	disagree
3	disagree	agree	disagree

Majority-rule cannot be applied to all three formulas, since the result would be inconsistent (the three biconditionals cannot all be false). But which basis among the three distinguished in (3) should be used for voting? We see that each gives a different result:

Basis used	Impact under voting		
	$p \leftrightarrow q$	$q \leftrightarrow r$	$p \leftrightarrow r$
$\{p \leftrightarrow q, q \leftrightarrow r\}$	disagree	disagree	agree
$\{p \leftrightarrow q, p \leftrightarrow r\}$	disagree	agree	disagree
$\{q \leftrightarrow r, p \leftrightarrow r\}$	agree	disagree	disagree

And the three experts might have cyclic preference-rankings for the three bases. (For discussion of cyclic preferences and voting dilemmas, see Johnson, 1998.) The dilemma can be resolved by expanding the set $\{p \leftrightarrow q, q \leftrightarrow r, p \leftrightarrow r\}$ to include the variables p, q, r, which would be a basis distinguished by its simplicity. But the three experts may decline to pronounce themselves on p, q, r, or no longer be around to do so.

Aggregating probabilities

These difficulties can be avoided if we are willing to embrace more nuanced beliefs. Then the data in (4) suggest assigning one-third probability to each of $p \leftrightarrow q$, $q \leftrightarrow r$, and $p \leftrightarrow r$. Such an assignment comes from interpreting an expert's opinion as probabilistic certitude (zero or unity), and averaging the numbers. The resulting probabilities have the virtue of *coherence* in the sense of consistency with the axioms of probability (it can be shown that averaging a set of coherent probability estimates of formulas yields a coherent estimate). But now the door is open to probabilistic opinions on the part of the experts (not just us), and averaging does not always work in this case, when we allow conditional probability estimates. Suppose, for example, that there were two experts with the following opinions:

(5)

Expert	q	$p \wedge q$	$p : q$
1	$.5$	$.1$	$.2$
2	$.25$	$.2$	$.8$

(We use *Prob* $(p : q)$ to denote the conditional probability of p given q.) The two sets of opinions in (5) are each probabilistically coherent, but their average — *Prob* $(q) = .375$, *Prob* $(p \wedge q) = .15$, and *Prob* $(p : q) = .5$ — is not coherent, since the ratio of *Prob* $(p \wedge q)$ to *Prob* (q) is $.4$ rather than the required $.5$.

Difficulties for aggregating probabilistic beliefs arise even in the absence of averaging. Suppose that one expert opines *Prob* $(p : q) = .6$, another *Prob* $(p) = .2$, and another *Prob* $(q) = .4$. Each expert has a distinct coherent opinion, and no averaging is necessary to adopt them all. Such simple aggregation is again unwise, however, since the union of the three judgments is incoherent.[2] A sensible aggregation policy must revise some of the opinions before adopting them.

The needed revision might be obtained through a protocol that facilitates communication and consensus among the three experts, convincing one or more to change their view. This is the method offered by the "Delphi System" and its variants (see Parenté & Anderson-Parenté, 1987; Cooke, 1991 for overviews and evaluation). The present paper explores the opposite technique of consulting each expert just once, and then minimally adjusting their probabilities "off line" so as to achieve coherence. This approach is useful in situations where experts are difficult to assemble. It is also convenient when they must estimate the probabilities of many events, both logically simple (like p and q) and complex (as in the conditional probability of p given q, or their conjunction, etc.). Checking coherence of the estimates may involve long computations thereby discouraging multiple cycles of revision and discussion.

Unifying disparate assessments of chance without further communication among experts generalizes problem (1), above, raised by Kornhauser & Sager (1986). The present paper offers one method for this kind of probabilistic aggregation. We picture a panel of experts, each offering estimates of confidence conceived as subjective probabilities. The matter is particularly challenging when different experts assign probabilities to events defined over distinct but overlapping sets of variables. In this case, we can distinguish three obstacles to unification.

(a) The set of probability estimates offered by a single expert may be incoherent, even apart from its juxtaposition with the estimates of other experts (*incoherence within experts*).[3]

(b) Different experts evaluating the same event might assign it different probabilities (*inconsistency among experts*).

(c) Different experts evaluating logically related events might assign probabilities that cannot all be coherently accepted, as illustrated earlier (*incoherence among experts*).

Our approach responds at once to the three obstacles by collecting all the judgments into one set, and then searching for minimal modifications that render the entire set of estimates probabilistically coherent. For an application of this technique, consider geopolitical forecasting (or intelligence assessment). Agents in different parts of the world may offer probabilities for simple and complex events over intersecting sets of variables (involving oil production, political stability, armament, and so forth). Communication among agents might be impractical for security reasons. A unified picture of the international scene could nonetheless be constructed by pooling the judgments and then adjusting them for coherence.

We describe below an efficient algorithm for probability aggregation and report experiments on its use. Of particular interest is the impact on the stochastic accuracy of group judgment before and after aggregation. Along the way, theorems are presented that serve as partial justification for our aggregation method. But in general, we eschew the axiomatic approach in favor of experimentally documenting a method that works.

Comparison to earlier studies

Our perspective differs from others' inasmuch as it allows different experts to estimate the probabilities of distinct events over sets of variables that overlap only partially. Thus, one expert might estimate the chances of $p \wedge q$ and of $q : r$, another might estimate r and $r \vee \neg s$, and so forth. In contrast, the usual goal is to aggregate opinions about the same event (e.g., Rowe, 1992; Ariely et al., 2000), or about the same partition of events. The major schemes for aggregation in the latter contexts involve Bayesian models (Morris, 1974; Clemen & Winkler, 1993; Clemen, Jones & Winkler, 1996) and linear averaging (Wagner 1984; Wallsten, Budescu, Erev & Diederich, 1997; see also Pennock, 1999, for an approach based on an analogy to securities markets). Literature reviews in this area include Clemen (1989; Clemen & Winkler, 1999), and Genest & Zidek, (1986). These approaches seem not to apply straightforwardly to the situation envisioned here, where events relate to each other in complex ways, are evaluated by partially overlapping sets of experts, and are plagued by intra- and inter-judge incoherence.[4]

A special case of the aggregation problem is to restore coherence to the estimates of a single individual. So far as we know, the first scheme for off-line adjustment of a single person's incoherent probabilities is due to de Finetti (1974). It is discussed below, and used as a comparison to our own approach. Another important scheme is offered in Lindley, Tversky & Brown (1979). Incoherence is there conceived as arising via error from an underlying source of coherent probabilities (not consciously accessible to the judge herself). The observer must infer the underlying coherent probabilities on the basis of a 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. Bayes' theorem then allows calculation of the most likely underlying assessments of chance given the stated ones. Various assumptions simplify the desired calculations, but they ultimately require nonlinear optimization. Since the required distributions are difficult to evaluate empirically, and still require a complex optimization step, it strikes us as simpler to forgo the former and proceed at once to the latter. Additionally, the justification for Lindley et al.'s approach does not extend to aggregating the estimates of a panel of experts. This is because the incoherence of a panel cannot be assumed to arise from underlying shared and coherent convictions (not everyone has the same opinions, even subconsciously). Our goal is therefore to develop an efficient algorithm that brings coherence to a body of probability estimates (due either to an individual or a panel) by modifying them in a minimal way. If the modification were not minimal, we might lose whatever insight exists in the original judgments; and the judge(s) may reject our amendments. The general idea of minimally deforming a set of beliefs to restore consistency is familiar from the copious literature on belief revision (see P. Gärdenfors, 1988; S. O. Hansson, 1999, and references cited there).

An alternative to fixing incoherent estimates by an individual judge is to "fix" the judge beforehand. For attempts to instill statistically mature thinking in naive judges, see Fong, Krantz & Nisbett (1986), Nisbett, Fong, Lehman & Cheng (1987), and Schaller, Asp, Rosell & Heim (1996). A related approach is to elicit probability estimates in a way that obviates incoherence, e.g, by calculating the coherent response set prior to each judgment. For discussion of elicitation techniques, see Alpert & Raiffa (1982), Holtzman & Breese (1986), von Winterfeldt & Edwards (1986), Henrion (1987), Morgan & Henrion (1990), and Klayman and Brown (1993). Druzdzel & van der Gaag (1995) describe techniques for computing coherent response sets, even for judgments involving independence and conditional independence. One drawback to constraining the judge's present estimate as a function of past estimates is that the final set of (coherent) estimates may depend on the order in which they are elicited.

We now review elements of subjective probability (developed more fully in Jeffrey, 1983, Nilsson, 1986, and Halpern, 2003, among other sources). Afterwards, the problem of restoring coherence to a corpus of judgments (drawn either from a single expert or a panel) is formulated in terms of optimization. Our algorithm for solving the optimization problem is then presented, followed by its evaluation in an experimental study. The theory described below bears exclusively on point estimates of probability rather than upper and lower bounds. Probability manipulations involving bounds are studied in Walley (1991, 1996), Biazzo & Gilio (2000), and references cited there. We also limit attention to the case of Boolean variables, i.e., taking no more than true and false (or 1 and 0) as values. We stress that there is no difficulty extending our framework and algorithm to random variables involving any finite set of values (e.g., alternative percentages of oil price decline in the next quarter). It simplifies the present discussion to stick with the Boolean case. Similarly, we rely on the inessential assumption that all experts in a given panel have equal credibility. As noted below, it is straightforward to allow different judges to influence the approximation algorithm as a function of their reputation for stochastic accuracy.

The standard picture of subjective probability

We start with a set of n variables, coding sentences with determinate (if unknown) truth-value (for example, *The Saudi dynasty collapses next year*). The variables give rise to a larger set of formulas defined in the usual way from the sentential connectives $\neg$ (negation), $\wedge$ (conjunction), and $\vee$ (disjunction).

The language is interpreted via mappings called *states*. Each state has the set of n variables as domain, and truth and falsity as range. A state thus represents a potential reality or "state-of-the-world." A state α *satisfies* a formula φ if α makes φ evaluate to true via standard truth-tables, denoted $\alpha \models \varphi$.

Let $\mathcal{A}$ be the set of 2^n states over the n variables. A *(probability) distribution* is a mapping $\mathit{Prob} : \mathcal{A} \to [0, 1]$ such that $\Sigma_{\alpha \in \mathcal{A}}\ \mathit{Prob}\,(\alpha) = 1$. Formulas inherit their probabilities from the states that satisfy them. That is, given a distribution Prob and a formula φ, $\mathit{Prob}\,(\varphi) = \Sigma_{\alpha \models \varphi}\ \mathit{Prob}\,(\alpha)$. The probability of a formula is thus the sum of the probabilities of the states that satisfy the formula.

When probabilities are assigned to a formula (or the event it encodes), we call the probability "absolute." Probabilities assigned to one event assuming the truth of another are called "conditional." Conditional probabilities are computed via the familiar ratio

$$\mathit{Prob}\,(\varphi : \psi) \;=\; \frac{\Sigma_{\alpha \models \varphi \text{ and } \alpha \models \psi}\ \mathit{Prob}\,(\alpha)}{\Sigma_{\alpha \models \psi}\ \mathit{Prob}\,(\alpha)} \;=\; \frac{\mathit{Prob}\,(\varphi \wedge \psi)}{\mathit{Prob}\,(\psi)},$$

provided that the denominator is positive.

Optimization problem

Now we return to the problem of amending an incoherent corpus of probability estimates. The estimates are conceived as arising from a panel of experts, whose opinions are pooled to form a single set. As a special case (noted earlier), the panel might consist of just a single expert. In this case, the problem reduces to amending an individual's judgment should it suffer from incoherence.

We represent the panel's estimates by two lists, one for absolute probabilities, the other for conditional. We put (φ, x) in the first list just in case some expert estimated the probability of formula φ to be x. We put (χ, ψ, y) in the second list just in case some judge estimated the conditional probability of χ given ψ to be y. It is possible for (φ, x) to occur twice in the list if different experts make this same estimate. Likewise, the list might include both (φ, x) and (φ, y) for $x \neq y$. Similar remarks apply to conditional probabilities. The lists are *coherent* if there is some probability distribution that implies them; otherwise they are *incoherent*. Note that by this definition, the pair of judgments $(\psi, 0), (\varphi, \psi, .1)$ is incoherent because no conditional probability can be based on a conditioning event of probability zero.

In the typical case, the overall list $(\varphi_1, x_1) \ldots (\varphi_k, x_k)$, $(\chi_1, \psi_1, y_1) \ldots (\chi_j, \psi_j, y_j)$ is incoherent, and we seek to reconstruct it via a close distribution. The resulting optimization problem can be stated as follows.

(6) OPTIMIZATION PROBLEM: Let lists $(\varphi_1, x_1) \ldots (\varphi_k, x_k)$ and $(\chi_1, \psi_1, y_1) \ldots (\chi_j, \psi_j, y_j)$ be given. Find a (coherent) distribution *Prob* such that:

$$\sum_{i \leq k} \mid x_i - \mathit{Prob}\,(\varphi_i) \mid \quad + \quad \sum_{i \leq j} \mid y_i - \mathit{Prob}\,(\chi_i : \psi_i) \mid$$

is minimized. We also require that $\mathit{Prob}\,(\psi_i) > 0$ for all conditioning events ψ_i from the second list that also appear as one of the φ_ℓ's of the first list.

The latter requirement rules out such anomalous solutions as $\mathit{Prob}\,(q) = 0, \mathit{Prob}\,(p : q) = .5$, involving division by zero. No other formulas are required to have nonzero probability.

Absolute deviation is the proximity measure in (6), but it is easy to restate the problem in terms of squared deviation (we return to this point later). In either version, the optimization problem can be converted into a method for testing the satisfiability of Boolean formulas. Specifically, a formula φ is satisfiable if and only if the claim $\mathit{Prob}\,(\varphi) = .5$ is coherent. It is widely believed that there is no polynomial-time test of satisfiability (see Homer & Selman, 2001, §6.5). Hence both (6) and its squared deviation version are intractable in the general case. Indeed, even structurally simple subproblems of coherence-checking have been shown to be intractable (see Georgakopoulos, Kavvadias & Papadimitriou, 1988). Worse yet, in some circumstances there may not even *be* a minimum distance between the input lists and a coherent approximation to them; the distance can approach a lower bound without hitting it.

(7) EXAMPLE: The estimates $(q, 0)$, $(p, q, .5)$ are incoherent because no distribution *Prob* assigns a positive number to the ratio *Prob* $(p \wedge q)/$*Prob* (q) and zero to *Prob* (q). These judgments are approximated to within ϵ by *Prob* $(p : q) = .5$, *Prob* $(q) = \epsilon$, for any $\epsilon > 0$. But they can not be *perfectly* approximated, since they are not coherent.

We must therefore interpret (6) in the sense of finding a coherent approximation that is as close as feasibly possible to the input. For simplicity here, absolute and conditional events are given equal weight.

It is important to distinguish (6) from the problem of checking whether a corpus of judgments is coherent, and also from the problem of computing lower and upper bounds on the probabilities that can be coherently attributed to a new formula outside the given corpus. For discussion of the optimization routines used to solve the latter problems (variants of linear programming), see Lad (1996), Biazzo, Gillio, Lukasiewicz & Sanfilippo (2001), and references cited there. In the present setting the typical corpus is incoherent and we seek a new set of judgments to replace it.

For absolute deviation, an optimal distribution need not be unique. For example, the incoherent judgments *Prob* $(p) = .3$, *Prob* $(\neg p) = .6$ are equally well approximated by either *Prob* $(p) = .4$, *Prob* $(\neg p) = .6$ or *Prob* $(p) = .3$, *Prob* $(\neg p) = .7$. No other approximation gets closer. Such multiplicity can be avoided by the use of squared rather than absolute deviation, since the approximation *Prob* $(p) = .35$, *Prob* $(\neg p) = .65$ is the unique best solution in the sense of squared deviation. (This is easily shown by differential calculus.)

The optimization technique we shall describe can be used *equally well* for either optimization problem (involving absolute or squared deviation). There turns out to be little empirical difference, however, between optimizing in terms of absolute versus squared deviation. Measuring deviation in these alternative ways leads to corrected estimates of chance that are typically close to each other, and that possess similar stochastic accuracy (measured by the “quadratic score,” as explained later). Likewise, our results are affected only slightly by use of other natural measures of proximity, for example, based on relative entropy (Cover & Thomas, 1991). In the interest of economy, we shall frame much of the ensuing discussion in terms of absolute deviation, presenting just one analysis with squared deviation for purposes of comparison. There are two reasons for favoring absolute deviation. First, absolute deviation is the more intuitive measure of accuracy, hence more likely to be understood and accepted by people whose estimates are being altered. Second, if attention is limited to absolute events, linear programming (LP) can be used to solve (6) but not its squared version. Indeed, for absolute events, LP's solution to (6) is best possible. LP therefore provides a useful benchmark for evaluating the accuracy of rival optimization techniques for (6). We note that LP has long been a popular method for carrying out probability calculations (see Hailperin, 1996, Chapters 1 - 3). Comparison of LP to the approximation procedure we propose below is thus facilitated by use of absolute deviation as common coin.

The use of LP to solve (6) is explained and illustrated in Batsell et al. (2002) and need not be repeated here. Our LP method is close to a technique advanced in Jaumard, Hansen and Poggi de Aragão (1991), where column generation is employed to apply LP to large probability problems. Linear programming does not embody a general solution to the problem of coherent approximation, however, since it is not applicable to judgments involving ratios of probabilities, notably, when estimates are given for conditional events (or when there are claims of independence or conditional independence, which also involve ratios). We therefore now present a more general approach to (6), and thus a more general approach to off-line reconstruction of incoherent judgments.

Simulated annealing over probability arrays

To hunt for a probability distribution that closely approximates an input corpus of estimates, we apply a well-known search technique to a novel class of data-structures called *probability arrays*. Probability arrays are sufficient to represent "sparse" probability distributions in a sense now made clear. We rely on the following observation, demonstrated in Batsell et al. (2002). [The observation is exploited in a similar way in Fagin et al. (1990). Its proof relies on a familiar fact about linear programming (Chvátal, 1983, Thm. 9.3).]

(8) FACT: Let formulas $\varphi_1 \cdots \varphi_k$, and pairs $(\chi_1, \psi_1) \cdots (\chi_j, \psi_j)$ of formulas be given. For every distribution *Prob* there is a distribution *Prob** such that:

  (a) *Prob** assigns positive probability to at most $k + j + 1$ truth assignments;

  (b) $\mathit{Prob}^*(\varphi_i) = \mathit{Prob}(\varphi_i)$ for every $1 \leq i \leq k$;

  (c) $\mathit{Prob}^*(\chi_i : \psi_i) = \mathit{Prob}(\chi_i : \psi_i)$ for every $1 \leq i \leq j$ with $\mathit{Prob}(\chi_i : \psi_i)$ defined.

It follows immediately from (8) that no distribution approximates the lists $(\varphi_1, x_1) \ldots (\varphi_k, x_k)$ and $(\chi_1, \psi_1, y_1) \ldots (\chi_j, \psi_j, y_j)$ of (6) better than some distribution that assigns positive probability to at most $k + j + 1$ states. This is true regardless of the measure that defines goodness of approximation [e.g., absolute deviation, as in (6), or squared deviation].

If we interview experts in a situation involving many variables, the number of states will dwarf the number of probability estimates. If there are 30 variables, for example, there are more than a billion states. So (8) tells us that an optimal coherent approximation may be found among the *sparse* distributions — those assigning positive probability to few states. The advantage to searching through the subspace of sparse distributions is that they can be compactly represented. We have experimented with several representations, including *algebraic decision diagrams* (Bahar et al., 1997; see Batsell et al., 2002). A simple structure turns out to work best. It is described next.

By a *probability array* let us mean a matrix of the following kind.

p	1	0	$*$	0
q	0	$*$	$*$	0
r	1	0	1	0
	.2	.1	.4	.3

Each row except the last represents a variable. The entries in the main part of the matrix represent truth, falsity, and "unspecified" — via 1, 0, and $*$. In this example, the first column in the main part stands for the state that makes p and r true, and q false. The second column represents *both* the state that makes all three variables false, and the state that makes just p and r false, q true. The third column stands for the set of four states in which r is true. The last row shows the probability mass shared by the set of states designated in a given column. The four states in the third column (**1) are thus each assigned $.1$. If a state shows up in more than one column, it receives the sum of the weights accruing to each. If a state shows up in no column, it receives zero weight. It is easy to see that probability arrays suffice to represent all sparse distributions in the sense of the fact shown next.

(9) FACT: Every distribution over n variables that assigns positive probability to no more than m states is represented by some probability array of dimension $(n + 1) \times m$.

In conjunction with (8), Fact (9) implies that the search for a coherent approximation to m probability estimates can be limited to probability arrays with $m + 1$ columns and $n + 1$ rows (where n is the number of variables). The estimates produced by the array are the closest possible coherent reconstruction of the original judgments. Indeed, with m columns for m positive states, we can even avoid the use of the "unspecified" symbol $*$ in probability arrays. Use of $*$'s, however, enriches the class of representable distributions and turns out to be helpful in our search procedure.

To search through probability arrays we have tried genetic algorithms (Michalewicz, 1992) but have had more success with simulated annealing (van Laarhoven, 1988).[5] Simulated annealing is an optimization procedure of wide application (Press, Teukolsky, Vettering & Flannery, 1992, §10.9). Here it is used to locate a probability array whose probabilities for a fixed, input corpus of (typically incoherent) estimates is minimized in the sense of (6). Details and variants of our approach are presented in Batsell et al. (2002). It suffices for the present discussion to indicate how random neighbors to a given probability array are generated in the annealing algorithm. With small probability (around $.1\%$), each non-bottom entry is replaced with a random choice among $\{0,\ 1,\ *\}$. Also with small probability, bottom entries are multiplied by a random choice between $.9$ and 1.1. The bottom row is then renormalized to sum to unity. The foregoing numbers were chosen empirically, after they were shown to yield rapid convergence to close approximations of input judgments.

Search in simulated annealing has a local character (only nearest neighbors are examined). To reach one probability array from another it is therefore necessary to find small steps that connect them. The presence of $*$ increases the chance of finding such a path by augmenting the number of distinct distributions accessible from any given probability array. To summarize:

> In what follows we use the term SAPA to denote the use of simulated annealing to search for a sparse probability distribution that approximates an input corpus of estimates, in the sense of (6). Sparse distributions are represented by probability arrays whose "neighbors" are defined by small perturbations of the matrix that composes the array. [More discussion of SAPA is available in Batsell et al. (2002).]

A principal virtue of SAPA is its computational feasibility on large problems. One test of the method involved sets of 200 simulated estimates of chance for complex events (both conditional and absolute) over 50 variables (more than a quadrillion states). SAPA reached close coherent approximations to these artificial judgments in little more than two hours on a Pentium III processor. Details of the test are available in Batsell et al. (2002). Note that linear programming cannot be easily applied to problems of this size, even if all events are absolute.

Because the optimization landscape might have infinitely deep holes [See Example (7), above], simulated annealing is not guaranteed to reach satisfactory solutions in every case. In practice, however, running the algorithm on the same problem from different random starting points reliably produced solutions in our computational experiments. The obtained solutions were similar but not identical, indicating the presence of local minima. It will be seen below that the approximations actually achieved are close to "best possible" when the latter criterion can be feasibly computed.

We summarize our strategy for aggregating the probability estimates of a panel of judges.

(10) AGGREGATION TECHNIQUE: Pool the estimates to form a single corpus of probability estimates, then apply SAPA to create a coherent set of probabilities that approximate all the original judgments at once. The latter probabilities are taken to be the group opinion.

Note that (10) allows each estimate to have equal influence on the output set of coherent probabilities. There is no difficulty in biasing SAPA to more closely approximate credible experts compared to others (credibility might be established, for example, by the accuracy of past predictions). It suffices to assign multiplicative coefficients to the judgments $(\varphi_i, x_i), (\chi_\ell, \psi_\ell, y_\ell)$ appearing in the optimization problem (6), with higher coefficients attached to judgments issuing from experts of greater authority. For simplicity, we here give every voice equal weight.[6]

Let us acknowledge the possibility that aggregating opinion via (10) may result in judgments that are congenial to nobody on the panel. In the simplest case, one expert might issue *Prob* $(p) = .9$, and another *Prob* $(p) = .1$. If aggregation yields *Prob* $(p) = .5$ then both may feel that their respective convictions have been misrepresented by an expression of ignorance. This risk is the consequence of aggregating without communication among experts. We defended such a strategy above, but here add that communication between two experts is not guaranteed to budge either of them. Indeed, social psychologists have provided ample documentation for opinion polarization after discussion of the same evidence presented to both parties! (See Lord, Ross & Lepper, 1979; Vallone, Ross & Lepper, 1985.)

Turning to other issues, let us note a plausible alternative to (10). It consists in first applying SAPA to each expert separately, thereby enforcing individual coherence before pooling the judgments of different experts. The resulting pool would once again be incoherent, and SAPA would be applied a second time. In contrast, (10) applies SAPA just once, to the pool of "raw" judgments without prior correction. We prefer (10) to the two-stage alternative because it allows the experts' opinions to interact with less interference from us. Moreover, experimental comparison of the two methods favors the one-stage (10); the stochastic accuracy of group judgment is better after applying (10) compared to the two-stage alternative. For brevity, we do not pursue the comparison in what follows.

Finally, to avoid misunderstanding, let us observe that (10) requires no commitment to the specific sparse distribution that provides close approximation to the input estimates of chance. The latter distribution is "thrown away" at the end of the process; only the corrected estimates are retained. Indeed, the input estimates will usually be approximated equally well by other distributions, including dense ones. We correct the input judgments using a probability array because Fact (9) guarantees that no other distribution gets closer to the original estimates. But we do not go on to accept the distribution as a model of uncertainty in the environment.

An experiment designed to test SAPA on predictions of economic indicators is now reported. All computation were carried out on Pentium 4 processors.

Experiment

Stimuli

We created 30 variables bearing on the United States' economic situation in the fourth quarter of 2001. Some concerned the performance of individual stocks; others concerned leading indicators. All 30 are listed in Appendix 1. In the last two weeks of September 2001, we asked people to estimate the probability of events based on these variables. Each participant was assigned 10 variables drawn randomly from the larger set of 30. Participants then estimated the probabilities of these 10 events plus the probabilities of 36 complex events built from them.

——- INSERT TABLE 1 ABOUT HERE ——-

The 36 complex events included an individually randomized selection of 6 events drawn from each of the categories shown in Table 1. The top row of the table represents the 10 variables chosen individually randomly for the judge. The second row represents the 6 conditional events constructed individually randomly from the 10 variables. Likewise, there were 6 conditionals with negated conditioning event, 6 conjunctions, 6 conjunctions with negated second conjunct, 6 disjunctions, and 6 disjunctions with negated second disjunct.

Participants and procedure

As judges, we recruited 21 graduate students in business, math, or statistics at Rice University. There were also 26 undergraduates taking statistics courses, for a total of 47 judges altogether.

Prior to collecting their estimates, participants received a short lecture on interpreting stock prices, the meaning of the economic indices used in the study, the inclusive meaning of *or*, and the conditional reading of the expression *assuming that*. It was made clear that all events were relative to the fourth quarter of 2001. Prizes were offered for stochastic accuracy as measured by the *quadratic penalty*, which was explained to participants (see below for the definition of this penalty).

Subsequent to these explanations, participants entered their 46 estimates of chance via a website that carried out all individual randomizations. Probabilities were represented numerically with the aid of a slider.[7] At the end of the fourth quarter of 2001, everyone's quadratic penalties were computed, and $500 in gift certificates were distributed to participants with lowest penalties.

Average estimates and incoherent responses

For each participant we calculated the average probability assigned to the 7 kinds of events they evaluated. The means of these averages are shown in Table 2. Correlated t-tests reveal that the mean estimate for variables (p) is reliably greater than for conjunctions ($p \wedge q$), and less than for disjunctions ($p \vee q$) [$t(46) = 7.26, 6.30$]. These differences are reasonable, since a conjunction tends to have smaller probability than its conjuncts, and disjunctions tend to have greater probabilities than their disjuncts. But it will now be seen that such global tendencies mask profoundly incoherent judgment.

——- INSERT TABLE 2 ABOUT HERE ——-

We assessed the coherence of each participant's estimates in terms of the four probability laws displayed in Table 3. For example, we considered law (a) to be violated whenever a given participant's probability for a conjunction $p \wedge q$ exceeded the probabilities of p or q, or fell short of the sum of the latter probabilities minus one. [Constraint (a) in Table 3 is stated in Suppes (1966). The other laws are easy corollaries of probability theory.] Since each participant estimated six conjunctive events along with all the variables that compose them, there were six opportunities to violate law (a). The other laws similarly yield six potential violations per participant.

——- INSERT TABLE 3 ABOUT HERE ——-

The first two columns of Table 4 show the average number of violations of each law. The 26 undergraduates in the sample averaged 15.46 violations of all four laws, out of a total of 24 possible violations (S.D. $= 4.98$). The 21 graduate students averaged 10.24 total violations (S.D. $= 6.06$), significantly less than the undergraduates (one way ANOVA, $F(1, 45) > 10.53$, $p < .002$). The relevant feature of the graduate student judgment is nonetheless the massive violation of probability laws. Only two of the 47 participants succeeded in avoiding any violation of the four laws in Table 4 (and even these two participants were incoherent by other measures).[8]

——- INSERT TABLE 4 ABOUT HERE ——-

The incoherence is numerically quite robust. The last columns in Table 4 show the average number of violations of the four probability laws that result from tolerating inconsistencies that can be erased by relaxing the laws in Table 3 by $.05$ or $.10$. For example, relaxing law (a) by $.05$ yields:

$$\boxed{\begin{array}{lcl} \mathit{Prob}\,(p \wedge q) & \leq & \min\{\mathit{Prob}\,(p), \mathit{Prob}\,(q)\} + .05 \\ \mathit{Prob}\,(p \wedge q) & \geq & \mathit{Prob}\,(p) + \mathit{Prob}\,(q) - 1.05 \end{array}}$$

Bar-Hillel (1973) argues that conjunctive events have a tendency towards overestimation in human judgment, whereas disjunctive events have a tendency towards underestimation. This predicts that most violations of (a) in Table 3 will be of the first inequality rather than the second, and similarly for constraints (b) - (d). Consistent with Bar-Hillel's prediction, the numbers of participants (out of 47) who violated the first inequalities in (a) - (d) more often than the second were 32, 35, 31, and 32, respectively ($p < .02$ by a binomial test).

Coherent approximation via SAPA

To render the estimates coherent, we applied SAPA separately to each participant's data set of 46 judgments (aggregation of opinion was not yet attempted). There were thus 47 optimizations, each involving 10 variables (namely, the 10 variables chosen randomly out of 30 for a given judge). For each participant we calculated the *mean absolute deviation*, or MAD, between her 46 probability estimates and those offered by the best approximating distribution discovered for her. Across the 47 participants, the average MAD was $.09$, as shown in the first line of Table 5. Each optimization took less than 20 seconds.

——- INSERT TABLE 5 ABOUT HERE ——-

Perfect approximation — a MAD of zero — is ruled out because of the incoherence that characterizes the estimates. For just the 34 estimates of absolute events, the closest possible approximation can be calculated using LP. When we applied it to the absolute judgments of each participant, the average MAD obtained was $.079$, as shown in the second line of the table. For comparison, we again applied SAPA to each participant, retaining only the 34 estimates of absolute probability. As shown in the third line, the average MAD this time was $.085$. Thus, for absolute events, SAPA comes within 8% of producing optimal coherent approximations.

Stochastic accuracy before and after coherent approximation

Two measures of accuracy

There is more to useful judgment than coherence. If the modifications imposed by SAPA rob the original estimates of their accuracy, the judge might prefer that we left her estimates in their original state. To find out whether our procedure degrades the knowledge inherent in a corpus of judgment, we need first to define the accuracy of a probability estimate. A standard measure of stochastic accuracy is the *quadratic penalty* (Brier, 1950; von Winterfield and Edwards, 1986) defined as follows.

(11) DEFINITION:

(a) Suppose that p is the estimated probability of an (absolute) event E. The *quadratic penalty* for the estimate is $(1-p)^2$ if E is true. It is p^2 if E is false.

(b) Suppose that p is the estimated probability of the conditional event "E assuming that F." The *quadratic penalty* for the estimate is $(1-p)^2$ if both F and E are true. It is p^2 if F is true but E is false. The quadratic penalty is undefined if F is false.

(c) Given a corpus of estimates for absolute and conditional events, the *quadratic penalty* for the corpus is the average of the quadratic penalties that are defined for events in the corpus.

Accurate judgment corresponds to low quadratic penalties. (Expected quadratic penalties are minimized by truthful reporting of subjective probability; see von Winterfield and Edwards, 1986). A contrasting measure of stochastic accuracy — for which accuracy is reflected in high scores — was also employed. Following Yates (1990, Ch. 3), we define the *slope* of a corpus of judgments to be the average probability assigned to events that come true minus the average probability assigned to events that do not come true. In this definition, a conditional event $(A : B)$ is considered to be undefined if B does not occur. Otherwise, it is considered to occur if and only if A occurs.

Quadratic penalties before and after coherent approximation

In this subsection and the next, we consider stochastic accuracy of individual subjects before and after rendering their estimates coherent via SAPA. The accuracy of aggregation is discussed in the next section.

At the end of the fourth quarter of 2001 we determined the truth values of all the events in the study, calculated everyone's quadratic penalty and handed out prizes accordingly. In the experiment, 19 of the 30 variables came true; 11 turned out to be false. The truth-values of all complex events in the experiment follow from the truth-values of the variables. As shown in line (a) at the top of Table 6, the average quadratic penalty achieved by the 46 judges in Experiment 1 was $.314$ (S.D. $= .077$). This penalty does not suggest much accuracy in judgment, a point to which we return below.

——- INSERT TABLE 6 ABOUT HERE ——-

We next calculated the quadratic penalty for each judge's estimates after transformation by SAPA. As shown in line (b) of Table 6, the average quadratic penalty drops to $.285$ (S.D. $= 076$). The improvement between the penalties for raw and transformed estimates is significant by correlated t-test [$t(46) = 7.6$, $p < .001$]. Forty of the 47 participants had lower quadratic penalty after application of SAPA compared to before ($p < .001$ by a binomial test assuming equal probability of improved versus worse penalties).

As discussed earlier, SAPA can be adapted to the squared-deviation version of the optimization problem (6). In line (c) of Table 6 we see that application of the squared-deviation version of SAPA to the judge's raw estimates lowers the average quadratic penalty even further, to $.277$ (not significantly lower than the penalty associated with the absolute-deviation version of SAPA). We prefer the use of absolute to squared deviation in formulating (6) because (a) absolute deviation is more easily interpreted by the experts whose opinions are transformed by SAPA, and (b) it facilitates comparison of our approximation results to linear programming (see above). There would seem to be nothing inappropriate about using absolute deviation as a measure of proximity to judgments while using quadratic penalty as a measure of proximity to truth; different things are compared. Moreover, we are equally interested in slope as a measure of stochastic accuracy, and this measure is linear rather than quadratic. Line (c) of Table 6 is nonetheless useful for suggesting the robustness of SAPA.

We now seek a benchmark in order to evaluate the improvements in quadratic penalty just discussed. Optimal improvement is not an interesting concept, since the quadratic penalty can be reduced to zero by setting each estimate to its truth-value (0 for false, 1 for true). If attention is restricted to absolute judgments, however, the following theorem of de Finetti (1974) provides an interesting comparison with our results.

(12) THEOREM: For every incoherent set I of probability estimates over absolute events (no

> conditional probabilities) there is a coherent set C of estimates over the same events such that the quadratic penalty for C is lower than the penalty for I no matter which state is true.

In other words, de Finetti showed that incoherent estimates over a set of *absolute* events can always be replaced by coherent estimates that have lower penalties in every state.

Theorem (12) can be extended to some sets of estimates involving conditional events; see, for example, Bernardo and Smith (1994, p. 89). The theorem cannot be extended, however, to *every* corpus of estimates because of examples like $\{(p,\ q,\ .5,), (q,\ 0)\}$. An easy argument shows that no coherent correction in this case lowers the quadratic score in all states.[9]

Returning to absolute events, de Finetti's (1974) proof of Theorem (12) describes a method for constructing coherent approximations to incoherent estimates of absolute events that lowers the quadratic penalty in all states. (It leaves coherent estimates untouched.) Following the discussion in Joyce (1998), we can describe de Finetti's coherent approximation as follows. Let (φ_i, x_i) be a sequence of n pairs consisting of an (absolute) event φ_i and an estimate x_i of its probability. Let a length n vector $\overrightarrow{v}$ of 0's and 1's be called "achievable" if there is some state A over the variables appearing in the φ_i's such that the ith coordinate of $\overrightarrow{v}$ is the truth-value of φ_i according to A. De Finetti showed that the estimates x_i are coherent if and only if the vector $\overrightarrow{x} = (x_1 \cdots x_n)$ is a weighted sum of all achievable vectors, where the weights are nonnegative and sum to unity. If the vector $\overrightarrow{x}$ is not coherent, de Finetti replaces it with the closest coherent vector $\overrightarrow{y}$ in the Euclidean sense, and shows that $\overrightarrow{y}$ has lower quadratic penalty than $\overrightarrow{x}$ in every state. The de Finetti point can be calculated via quadratic programming (Luenberer, 1984) using routines in the MATLAB environment. In what follows, we use the symbol FINETTI to denote the method consisting of constructing such a $\overrightarrow{y}$ for incoherent $\overrightarrow{x}$.

Note that even for absolute events, FINETTI is not guaranteed to deliver lower quadratic penalties than a rival method like SAPA. Theorem (12) only guarantees lower quadratic penalties than the incoherent target.

For each of our participants, we applied FINETTI to her subset of 34 absolute estimates (the 12 conditional events appearing in Table 2 are excluded from the present analysis). As shown in line (d) of Table 6, the average quadratic penalty for absolute events over all participants was .309 (S.D. = .079). Applying FINETTI to the absolute estimates lowers the average quadratic penalty to .272 (S.D. = .074); see line (e) of Table 6. All 47 participants have lower quadratic penalty after approximation by FINETTI [this is guaranteed by Theorem (12) and the fact that all subjects were incoherent]. For comparison to SAPA, for each participant, we calculated the quadratic penalty for SAPA's coherent approximations of just the 34 absolute estimates [line (f) of Table 6]. The average quadratic penalty of the coherent approximations was $.276$ (S.D. $=$ 081), thus only slightly higher than with FINETTI. The quadratic penalty for absolute events was better using SAPA compared to FINETTI for 24 of the 47 participants. SAPA is thus comparable to FINETTI for absolute events. Of course, SAPA can also be applied to conditional events, which is not the case for FINETTI. Additionally, SAPA can be applied in the context of many variables whereas the quadratic programming required for FINETTI limits it to small problems (each participant in our study confronts a set of just 10 variables).

Slope before and after coherent approximation

One noteworthy feature of our data is that participants showed little insight into the upcoming fourth quarter of 2001. Had they answered with probability one half to every question, they would have been incoherent but their penalty would have been $.25$, lower than what they typically achieved. Perhaps we can forgive them for not being able to predict events in the turbulent autumn of 2001, right after the events of September 11, and right before the collapse of Enron. Note that one of the 30 variables stated that the stock price of Enron corporation would increase by more than 10% in the fourth quarter (see Appendix 1). This seemed plausible at the time.

In any event, the participants' judgment looks better if we change the criterion of accuracy to slope, whose natural threshold of ignorance is zero. (Recall that higher slopes signify greater accuracy, the reverse of quadratic penalty.) As seen in line (a) of Table 7, the average slope for the participants' raw estimates is $.053$ (S.D. $= .130$) which is reliably greater than zero [$t(46) = 2.81$, $p < .01$]. So by this measure, our judges seemed to have some inkling about the fourth quarter of 2001. As seen in line (b) of Table 7, greater insight comes from SAPA's coherent approximation of their estimates. The average slope of the reconstructed estimates is $.115$ (S.D. $= .129$). The difference between the two sets of slopes is reliable by correlated t test [$t(46) = 8.46$, $p < 001$]. Greater slopes for reconstructed compared to raw estimates appear for 40 of the 47 judges ($p < .01$ by binomial test).

Unfortunately, we cannot adapt FINETTI to a method that provides uniform improvement in slope for absolute judgments. That is, there is no analogue to Theorem (12) for slope. Like quadratic penalty, the slope of a set of estimates depends on which state describes the world; we say that slope is "relative to" states. The impossibility of uniform improvement in slope thus signifies the existence of an incoherent corpus of absolute judgments such that every coherent alternative has lower slope relative to some state. The matter is summarized in the following proposition, proved in Appendix 2.

(13) PROPOSITION: There are sets of incoherent estimates of probability for absolute events with the following property. For every coherent set of estimates for these events there is a state such that the slope of the incoherent estimates exceeds the slope of the coherent estimates relative to that state.

The present section has documented improvement in stochastic accuracy when incoherent estimates of chance are replaced by the coherent approximation provided by SAPA. The same phenomenon appears in three additional experiments of similar design (involving sports and stock prediction), discussed in Batsell et al. (2002) and in Deines et al. (2003). Overall, the four studies leave little doubt that coherent approximation has a tendency to improve stochastic accuracy. This phenomenon is similar to "bootstrapping" when predicting college grades or other quantitative variables: a linear model of the judge's estimates often predicts the variable better than the judge herself (Dawes, 1979; Dawes & Corrigan, 1974; Camerer, 1981). Rectifying probabilities has a feature not present in the context of linear models, however. Whereas there are no normative grounds for using a linear equation to predict college grades from SATs, high school grades, etc., there are persuasive reasons to prefer one's probabilities to be coherent. For example, 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. For discussion, see Resnik (1987), Coletti, Gilio & Scozzafava (1993), Gustason (1994), and Osherson (1995).

Aggregation of estimates

We now evaluate Strategy (10), discussed above, to aggregate the estimates of the 47 judges. Our focus is again stochastic accuracy. Does aggregation lead to superior estimates?

We pooled all estimates of chance offered by any of the 47 participants in our study of the fourth quarter of 2001. The resulting set of judgments may be called the *aggregate judge*. Recall that each participant made estimates for events defined from 10 variables, randomly chosen for him or her from a stock of 30. The aggregate judge therefore works in a probability space built from 30 variables. SAPA was used to find a close approximation in this space to all 2162 judgments at once. The computation required about an hour.

The mean absolute deviation of the computed approximation to the raw estimates was $.193$. This is double the deviation seen when coherent approximations were computed for one subject at a time (which averaged $.09$). Coherent approximation of numerous raw estimates seems destined to be looser than approximation of just a few judgments, since there are more conflicting demands on the approximation.[10] For these data, we could not compare our performance to LP (on absolute judgments) because there are more than a billion states. (For the same reason, FINETTI could not be applied.)

For each of the 47 judges, we compared the quadratic penalties arising from:

(i) the judge's original estimates;

(ii) the judge's estimates after revision by individual application of SAPA; and

(iii) the judge's estimates after revision in the aggregate, as in (10).

For (iii), the judge's estimate of the probability of an event φ was replaced by the estimate of φ obtained by applying SAPA to the aggregate judge (and similarly for conditional events).

The average quadratic penalty for SAPA applied to the aggregate judge is $.254$, as seen in line (g) of Table 6. This is lower than the average penalties for raw estimates and SAPA applied individually; see lines (a), and (b) of Table 6. The differences are reliable by correlated t test [$t(46) = 5.84, 3.15$, respectively]. For 32 out of the 47 subjects, the penalty is lower if their estimates were corrected as part of the group compared to correcting them individually ($p < .01$ by a binomial test). So it seems that coherent approximation squeezes more information out of human judgment if opinion is pooled rather than left atomized.

Fifteen pooled judges seem to be enough to get all the benefit of aggregation, at least in these data. When random subsets of 15 are drawn from our pool of 47 participants, we observe lower quadratic penalties for the aggregate compared to the 15 individuals. Beyond 15, further improvement for the aggregate judge is not detectable. This finding is consistent with other experimental studies on quantitative estimates; small panels suffice to obtain most of the benefit of aggregated judgment (see Ashton & Ashton, 1985, and references cited there).

The foregoing results concern quadratic penalty. When slope is used as a measure of accuracy there is only small benefit of aggregation. As shown in Table 8, line (c), the average slope after application of SAPA to the aggregate judge is $.123$. This is superior to slopes from raw estimates and SAPA applied individually, but the difference is reliable by correlated t-test only in the comparison to raw estimates [$t(46) = 4.37$]. Thirty-six of the 47 participants have higher slopes after SAPA applied in the aggregate compared to their raw estimates ($p < .001$ by binomial test). Only 26 of 47 have higher slopes in the aggregate compared to individual application of SAPA (not significant).

Even stronger affects of aggregation are seen in the three additional studies reported in Batsell et al. (2002).

Discussion

What explains the improvement in stochastic accuracy following coherent approximation by SAPA? Regarding quadratic penalty, suppose that the judge offers $(p,\ .6)$ and $(p \wedge q,\ .8)$. A closest coherent approximation revises both estimates to $.7$, which lies between the two original estimates. Since quadratic penalty is a convex function, the penalty is guaranteed to be lower for the two copies of $.7$ compared to $.6$ and $.8$. The foregoing explanation does not extend to improvements seen in slope. First, Proposition (13) shows that Theorem (12) has no analogue for slope. Moreover, slope is not a convex function of estimates.

Improvement in stochastic accuracy when opinion is aggregated presumably results from reduction of error variance through averaging. A rigorous formulation of this phenomenon (which we do not attempt here) is complicated by (a) the presence of incoherence within and between judges and (b) variability in the events evaluated by different judges. It might also be possible to understand the improvement in terms of some variant of Condorcet's Jury Theorem (Grofman, Owen & Feld, 1983). The "jurors" in our context, however, are not voting in a binary way but rather offering probabilities. Moreover, their opinions are likely to be correlated via reliance on information of the same provenance; such a situation is known to compromise the uniform tendency of well-informed group opinion to converge to the truth (Ladha, 1995; see

Berend & Paroush, 1998, for a different analysis of Condorcet-like phenomena).

It is worth observing that aggregation via Strategy (10) potentially allows every judge to influence the coherent probabilities constructed for every other judge. To illustrate, consider three judges, the first of whom faces events over the variables p, q, the second over q, r, and the third over r, s. Suppose that they offer the following estimates.

(14)

Judge 1:	$Prob\,(p \wedge q)$	$\approx$	1
Judge 2:	$Prob\,(r : q)$	$\approx$	1
Judge 3:	$Prob\,(s : r)$	$\approx$	1

Suppose also that Judge 3 offers an estimate of the probability of the variable s. Then, if the estimates (14) are retained in the aggregate, coherence requires that $Prob\,(s) \approx 1$. There is no such constraint on the probability of s in the absence of the estimate offered by Judge 1. Thus, in the aggregate, the estimates offered by Judge 1 influence the approximation of the estimates of Judge 3 even though the two judges face events over disjoint sets of variables. The improvement due to aggregation might result partially from such long-range influences, since without them aggregation reduces to correcting experts taken singly or in small groups (and it was seen that aggregation produced lower quadratic penalties than individual correction of estimates).

A potential alternative to the aggregation strategy (10) is to average the probability distributions that emerge from use of SAPA on individual participants. Such "linear pooling" of distributions has been analyzed in several axiomatic studies (see Wagner, 1984; Genest & Zidek, 1986). Observe, however, that linear pooling cannot be applied to the data of our experiment, since different participants evaluated events over distinct sets of variables (each was randomly assigned 10 variables from a starting set of 30). This obstacle to linear pooling would be typical in the kind of application discussed above (involving judges with overlapping expertise).

Finally, we note that practical problems of aggregation often involve "hard" constraints in the form of (coherent) probability estimates considered reliable, e.g., originating actuarially or in scientific models. Similarly, the causal structure of the environment may impose conditional independence among subsets of variables. (For background, see Castillo, Gutiérrez & Hadi, 1997.) Finding the best coherent approximation to expert judgment then becomes a problem of *constrained* optimization, and is more difficult to solve (Fletcher, 1986). A variant of SAPA can be adapted for this purpose (Kulkarni, Osherson, Predd & Vardi, forthcoming).

Notes

[1]It is generalized in List & Pettit (2002). Other studies of aggregation within the same framework have produced remarkable findings. See List (2003) and especially Dietrich & List (2004). Discussion of the repercussions of such findings for political theory is offered in Pettit (2001).

[2]Since $\mathit{Prob}\,(p) = .2$, $\mathit{Prob}\,(p \wedge q) \leq .2$. Inasmuch as $\mathit{Prob}\,(q) = .4$, the ratio of $\mathit{Prob}\,(p \wedge q)$ to $\mathit{Prob}\,(q)$ is at most $.5$, contradicting the opinion $\mathit{Prob}\,(p : q) = .6$.

[3]Incoherence is commonly observed when people are asked to estimate the probabilities of complex or conditional events; experts are not immune. See Baron (2000), Bonini, Tentori & Osherson (2004), Hastie & Dawes (2001), Sides, Osherson, Bonini & Viale (2002), and Yates (1990) for an entry to the literature on probabilistic incoherence. Classic papers in the field are collected in Kahneman & Tversky (2000) and Gilovich, Griffin & Kahneman (2002). For discussion of the coherence problem in the construction of Bayesian networks based on expert judgment, see van der Gaag et al. (1999).

[4]In this context, see Howard's (1989) discussion of incoherent probability estimates appearing within influence diagrams; the method proposed there, however, seems not to be general, and its outcome depends on the order in which sets of judgments are renormalized.

[5]Genetic algorithms conceptualize the search space as a set of formal genomes. Their "fitness" as solutions to an optimization problem determine the probabilities that given genomes will be preserved in the next "generation" of genomes to be evaluated, as well as the probability that they will participate in "crossover" with other genomes to generate new potential solutions. (Another source of new genomes is "mutation.") Mitchell (1996) provides an introduction to genetic algorithms, which have proven effective in a wide variety of optimization contexts. Their mathematical analysis is considered in Vose (1999). Simulated annealing is a gradient descent technique that incorporates stochastic movements uphill (more in early stages) to escape local minima (Kirkpatrick, Gelatt & Vecchi, 1983). After extensive use of genetic algorithms to search for a probability array that approximates a target set of judgments, we discovered that simulated annealing works faster by orders of magnitude, and finds probability arrays that are closer to the target.

[6]Alternative schemes for weighting judges are reviewed in Ferrell (1985). Application of SAPA to real decision-making should also be sensitive to correlation in the opinions of different judges (Hogarth, 1978). The practical consequences of ignoring both past reliability and inter-judge correlation remain unclear, however. Empirical results (e.g., Ashton & Ashton, 1985) as well as simulation studies (Johnson, D. Budescu & T. Wallsten, 2001) suggest that simple averaging is a robust strategy in the sense that the accuracy of the average probability is not extremely sensitive to violations of conditional pairwise independence of the judges (under certain reasonable conditions).

[7]The judge positioned a slider with endpoints labelled by 0 and 1. The position of the bar was expressed numerically in a field above the slider, thus providing the estimated probability. This method was chosen as a compromise between direct estimation of point probabilities and indirect estimation using spinners or other chance set-ups to create equivalent bets. The two methods do not always yield identical estimates of chance for the same events (Winkler, 1967).

[8]Specifically, each violated the constraint $\mathit{Prob}\,(p : q) = \mathit{Prob}\,(p \wedge q)/\mathit{Prob}\,(q)$ at least once. Overall, the judges violated the latter constraint more than half the time that it could be assessed. It could only be assessed when a given participant estimated the chance of $p \wedge q$ and the conditional probability of p given q. (She always estimated the chance of q.) Winkler (1971, p. 677) also reports surprisingly elementary violations of probability laws by naive judges asked to estimate the chances of point-spreads in upcoming football games. See also Wright & Ayton (1987, p. 91) for other violations of elementary laws concerning conjunction. In more general terms, a variety of unperceived influences on judgment have been experimentally demonstrated in recent years, e.g., noninformative numerical "anchors" enountered prior to estimation of a quantity (see Hastie & Dawes, 2001, Ch. 5).

[9]Coherence requires $\mathit{Prob}\,(q) = \epsilon > 0$, since otherwise $\mathit{Prob}\,(p : q)$ is undefined. Every state in which q is false yields lower quadratic penalty for the incoherent $\{(p,\ q,\ .5), (q,\ 0)\}$ compared to the coherent $\{(p,\ q,\ x),\ (q,\ \epsilon)\}$, for all $x \in [0, 1]$.

[10]Example: If the corpus of judgment is limited to $\{\mathit{Prob}\,(p) = .3, \mathit{Prob}\,(q) = .3\}$ then coherent "correction" modifies neither. If $\mathit{Prob}\,(p \wedge q) = .5$ is added, however, the three opinions cannot be accepted without change to some of them. More dramatic average change is necessary if $\mathit{Prob}\,(p \vee q) = .1$ is also added. Incompatible opinions of this character seem more likely when

pooling across experts rather than approximating just one. It is not easy to generalize the present example, however, since much depends on the logical forms of the judgments aggregated, and the correlation of belief across experts.

References

ALPERT, M., AND H. RAIFFA (1982): "A progress report on the training of probability assessors," in *Judgment under uncertainty: Heuristics and biases*, ed. by D. Kahneman, P. Slovic, and A. Tversky, pp. 294–305. Cambridge University Press, New York NY.

ARIELY, D., W. T. AU, R. H. BENDER, D. V. BUDESCU, C. B. DIETZ, H. GU, T. S. WALLSTEN, AND G. ZAUBERMAN (2000): "The Effects of Averaging Subjective Probability Estimates Between and Within Judges," *Journal of Experimental Psychology: Applied*, 6(2), 130–147.

ASHTON, A. H., AND R. H. ASHTON (1985a): "Aggregating subjective forecasts: some empirical results," *Management Science*, 31, 1499–1508.

——— (1985b): "Aggregating subjective forecasts: Some empirical results," *Management Science*, 31(12), 1499–1508.

BAHAR, R., E. FROHM, C. GAONA, G. HACHTEL, E. MACII, A. PARDO, AND F. SOMENZI (1997): "Algebraic decision diagrams and their applications," *Journal of Formal Methods in Systems Design*, 10(2/3), 171–206.

BAR-HILLEL, M. (1973): "On the subjective probability of compound events," *Organizational Behavior and Human Performance*, 9, 396–406.

BARON, J. (2000): *Thinking and Deciding (Third Edition)*. Cambridge University Press, Cambridge UK.

BATSELL, R., L. BRENNER, D. OSHERSON, S. TSAVACHIDIS, AND M. Y. VARDI (2002): "Eliminating Incoherence from Subjective Estimates of Chance," in *Proceedings of the 8th International Conference on the Principles of Knowledge Representation and Reasoning (KR 2002)*, pp. 353 – 364.

BEREND, D., AND J. PAROUSH (1998): "When is Condorcet's Jury Theorem valid?," *Social Choice and Welfare*, 15, 481 – 488.

BERNARDO, J., AND A. SMITH (1994): *Bayesian Theory*. John Wiley and Sons, New York.

BIAZZO, V., A. GILIO, T. LUKASIEWICZ, AND G. SANFILIPPO (2001): "Probabilistic Logic under Coherence: Complexity and Algorithms," in *2nd International Symposium on Imprecise Probabilities and Their Applications*, Ithaca NY.

BIAZZO, V., AND G. GILIO (2000): "A generalization of the fundamental theorem of de Finetti for imprecise conditional probability assessments," *International Journal of Approximate Reasoning*, 24, 251 – 272.

BONINI, N., K. TENTORI, AND D. OSHERSON (2004): "A new conjunction fallacy," *Mind & Language*.

BRIER, G. (1950): "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, 78, 1 – 3.

CAMERER, C. F. (1981): "General conditions for the success of bootstrapping models," *Organizational Behavior and Human Performance*, 27, 411 – 422.

CASTILLO, E., J. GUTIÉRREZ, AND A. HADI (1997): *Expert systems and probabilistic network models*. Springer, New York.

CHVÁTAL, V. (1983): *Linear Programming*. W. H. Freeman, San Francisco CA.

CLEMEN, R., AND R. WINKLER (1999): "Combining probability distributions from experts in risk analysis," *Risk Analysis*, 19, 187 – 203.

CLEMEN, R. T. (1989): "Combining forecasts: A review and annotated bibliography," *International Journal of Forecasting*, 5, 559–583.

CLEMEN, R. T., S. K. JONES, AND R. L. WINKLER (1996): "Aggregating Forecasts: An Empirical Evaluation of Some Bayesian Methods," in *Bayesian Analysis in Statistics and Economics*, ed. by K. M. C. D. A Berry, and J. K. Geweke. John Wiley & Sons, New York NY.

CLEMEN, R. T., AND R. L. WINKLER (1993): "Aggregating Point Estimates: A Flexible Modeling Approach," *Management Science*, 39(4), 501 – 515.

COOKE, R. (1991): *Experts in uncertainty*. New York: Oxford University Press.

COVER, T., AND J. THOMAS (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 B. CORRIGAN (1974): "Linear models in decision making," *Psychological Bulletin*, 81, 97–106.

DE FINETTI, B. (1974): *Theory of Probability, vol. 1*. John Wiley and Sons, New York NY.

DEINES, J., D. OSHERSON, J. THOMPSON, S. TSAVACHIDIS, AND M. Y. VARDI (2003): "Removing Incoherence from Subjective Estimates of Chance," Discussion paper, Rice University, Available at http://www.princeton.edu/∼osherson/Papers/cohere.pdf.

DIETRICH, F., AND C. LIST (2004): "Strategy-proof judgment aggregation," manuscript.

DRUZDZEL, M. J., AND L. C. VAN DER GAAG (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.

FAGIN, R., J. HALPERN, AND N. MEGIDDO (1990): "A Logic for Reasoning about Probabilities," *Information and Computation*, 87, 78 – 128.

FERRELL, W. R. (1985): "Combining Individual Judgments," in *Behavioral Decision Making*, ed. by G. Wright, pp. 111–145. Plenum Press, New York NY.

FLETCHER, R. (1986): *Pratical Methods of Optimization*. Wiley.

FONG, G. T., D. H. KRANTZ, AND R. E. NISBETT (1986): "The effects of statistical training about everyday problems," *Cognitive Psychology*, 18, 253–292.

GÄRDENFORS, P. (1988): *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. MIT Press, Cambridge MA.

GENEST, C., AND J. ZIDEK (1986): "Combining Probability Distributions: A Critique and an Annotated Bibliography," *Statistical Science*, 1(1), 114–135.

GEORGAKOPOULOS, G., D. KAVVADIAS, AND C. PAPADIMITRIOU (1988): "Probabilistic Satisfiability," *Journal of Complexity*, 4, 1–11.

GILOVICH, T., D. GRIFFIN, AND D. KAHNEMAN (eds.) (2002): *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge University Press, New York NY.

GROFMAN, B. G., S. L. OWEN, AND S. L. FELD (1983): "Thirteen theorems in search of the truth," *Theory and Decision*, 15, 261 – 278.

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

HAILPERIN, T. (1996): *Sentential Probability Logic*. Lehigh University Press, Bethlehem PA.

HALPERN, J. Y. (2003): *Reasoning about Uncertainty*. MIT Press, Cambridge MA.

HANSSON, S. O. (1999): *A Textbook of Belief Dynamics: Theory Change and Database Updating*. Kluwer Academic Publishers, Dordrecht.

HASTIE, R., AND R. M. DAWES (2001): *Rational Choice in an Uncertain World: The Psychology of Judgment and Decision Making*. Sage Publications, Thousand Oaks CA.

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

HOGARTH, R. (1978): "A Note on Aggregating Opinions," *Organizational Behavior and Human Performance*, 21, 42–43.

HOLTZMAN, S., AND J. BREESE (1986): "Exact reasoning about uncertainty: On the design of expert systems for decision support," in *Uncertainty and Artificial Intelligence*, ed. by J. F. Lemmer, and L. N. Kanal, pp. 339–346. North-Holland/Elsevier, New York NY.

HOMER, S., AND A. L. SELMAN (2001): *Computability and Complexity Theory*. Springer, New York NY.

JAUMARD, B., P. HANSEN, AND M. P. DE ARAGAO (1991): "Column Generation Methods for Probabilistic Logic," *ORSA Journal of Computing*, 3(2), 135–148.

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

JOHNSON, P. E. (1998): *Social Choice: Theory and Research*. Sage Publications, New York NY.

JOHNSON, T., D. BUDESCU, AND T. WALLSTEN (2001): "Averaging Probability Judgments: Monte Carlo Analyses of Asymptotic Diagnostic Value," *Journal of Behavioral Decision Making*, 14, 123–140.

JOYCE, J. M. (1998): "A Nonpragmatic Vindication of Probabilism," *Philosophy of Science*, 65, 575 – 603.

KAHNEMAN, D., AND A. TVERSKY (eds.) (2000): *Choices, values, and frames*. Cambridge University Press, New York NY.

KELLY, J. S. (1978): *Arrow Impossibility Theorems*. Academic Press, New York NY.

KIRKPATRICK, S., C. D. GELATT, AND M. P. VECCHI (1983): "Optimization by simulated annealing," *Science*, 220, 671 – 680.

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

KORNHAUSER, L. A., AND L. G. SAGER (1986): "Unpacking the Court," *Yale Law Journal*, 96(1), 82 – 117.

KULKARNI, S., D. OSHERSON, J. PREDD, AND M. Y. VARDI (forthcoming): "Restoring probabilistic coherence to estimates of probability via constrained optimization," Discussion paper, Princeton University.

LAD, F. (1996): *Operational subjective statistical methods*. Wiley & Sons, New York NY.

LADHA, K. K. (1995): "Information polling through majority rule voting: Condorcet's Jury Theorem with correlated votes," *Journal of Economic Behavior and Organization*, 26, 353 – 372.

LINDLEY, D. V., A. TVERSKY, AND R. V. BROWN (1979): "On the Reconciliation of Probability Assessments," *Journal of the Royal Statistical Society A*, 142 (Part 2), 146–180.

LIST, C. (2003): "A Possibility Theorem on Aggregation over Multiple Interconnected Propositions," *Mathematical Social Sciences*, 45(1), 1–13.

LIST, C., AND P. PETTIT (2002): "Aggregating Sets of Judgments: An Impossibility Result," *Economics and Philosophy*, 18.

LORD, C. G., L. ROSS, AND M. LEPPER (1979): "Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence," *Journal of Personality and Social Psychology*, 37, 2098 – 2109.

LUENBERGER, D. G. (1984): *Linear and Nonlinear Programming (2nd Edition)*. Addison-Wesley, Reading MA.

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

MORRIS, P. (1974): "Decision analysis expert use," *Management Science*, 20, 1233 – 1241.

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

NISBETT, R. E., G. T. FONG, D. R. LEHMAN, AND P. W. CHENG (1987): "Teaching reasoning," *Science*, 238, 625–631.

OSHERSON, D. (1995): "Probability judgment," in *Thinking*, ed. by E. E. Smith, and D. Osherson, pp. 35–76. MIT Press, Cambridge MA.

OSHERSON, D., D. LANE, P. HARTLEY, AND R. BATSELL (2001): "Coherent probability from incoherent judgment," *Journal of Applied Psychology*.

PARENTÉ, F. J., AND J. K. ANDERSON-PARENTÉ (1987): "Delphi Inquiry Systems," in *Judgmental Forecasting*, ed. by G. Wright, and P. Ayton, pp. 129–156. John Wiley & Sons, New York NY.

PENNOCK, D. M. (1999): "Aggregating Probabilistic Beliefs: Market Mechanisms and Graphical Representations," Ph.D. thesis, University of Michigan.

PETTIT, P. (2001): "Deliberative Democracy and the Discursive Dilemma," *Philosophical Issues*, 11, 268 – 299.

PRESS, W. H., S. A. TEUKOLSKY, W. T. VETTERLING, AND B. P. FLANNERY (1992): *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge UK.

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 *Expertise and Decision Support*, ed. by G. Wright, and P. F. Bolger, pp. 155–180. Plenum Press, New York NY.

RUSSELL, S. J., AND P. NORVIG (1995): *Artificial Intelligence: A Modern Approach*. Prentice Hall, Upper Saddle River NJ.

SCHALLER, M., C. H. ASP, M. C. ROSELL, AND S. J. HEIM (1996): "Training in statistical reasoning inhibits formation of erroneous group stereotypes," *Personality and Social Psychology Bulletin*, 22, 829–844.

SIDES, A., D. OSHERSON, N. BONINI, AND R. VIALE (2002): "On the reality of the conjunction fallacy," *Memory & Cognition*, 30(2), 191–198.

SUPPES, P. (1966): "Probabilistic inference and the concept of total evidence," in *Aspects of Inductive Logic*, ed. by J. Hintikka, and P. Suppes, pp. 49 – 65. North-Holland, Amsterdam.

VALLONE, R., L. ROSS, AND M. LEPPER (1985): "The hostile media phenomenon: Biased perception and perceptions of media bias in coverage of the Beirut massacre," *Journal of Personality and Social Psychology*, 49, 577 – 585.

VAN DER GAAG, L., S. RENOOIJ, C. WITTEMAN, B. ALEMAN, AND B. TAAL (1999): "How to Elicit Many Probabilities," in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, ed. by K. B. Laskey, and H. Prade, pp. 647–654, San Francisco. Morgan Kaufmann Publishers.

VAN LAARHOVEN, P. (1988): *Theoretical and computational aspects of simulated annealing*. Center for Mathematics and Computer Science, Amsterdam.

VON WINTERFELDT, D., AND W. EDWARDS (1986): *Decision analysis and behavioral research*. Cambridge University Press, New York NY.

VOSE, M. D. (1999): *The Simple Genetic Algorithm: Foundations and Theory*. MIT Press, Cambridge MA.

WAGNER, C. (1984): “Aggregating Subjective Probabilities: Some Limitative Theorems,” *Notre Dame Journal of Formal Logic*, 25(3).

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.

WALLSTEN, T., D. BUDESCU, I. EREV, AND A. DIEDERICH (1997): “Evaluating and combining subjective probability estimates,” *Journal of Behavioral Decision Making*, 10, 243–268.

WINKLER, R. L. (1967): “The assessment of prior distributions in Bayesian analysis,” *Journal of the American Statistical Association*, 62, 776–800.

——— (1971): “Probabilistic Prediction: Some Experimental Results,” *Journal of the American Statistical Association*, 66, 675–685.

WRIGHT, G., AND P. AYTON (1987): “The Psychology of Forecasting,” in *Judgmental Forecasting*, ed. by G. Wright, and P. Ayton, pp. 83–105. John Wiley & Sons, New York NY.

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

Appendix 1: The 30 variables used in the Finance experiment

1. The Standard & Poor's 500 Index increases.
2. The Standard & Poor's 500 Index outperforms the NASDAQ Composite Index.
3. The NASDAQ Composite Index increases.
4. General Electric's stock price increases.
5. Reliant Energy's stock price increases.
6. Exxon Mobil's stock price increases.
7. Enron's stock price outperforms Reliant Energy's stock price.
8. El Paso Corp's stock price increases.
9. Enron's stock price increases by greater than 10%.
10. Wal-Mart's stock price increases.
11. Amazon.com's stock price increases.
12. Sears Roebuck's stock price increases.
13. Wal-Mart's stock price outperforms Amazon.com's stock price.
14. The U.S. prime lending rate increases.
15. The price of crude oil decreases by more than 10%.
16. U.S. 30-year fixed mortgage rates decrease.
17. The U.S. retail sales rate increases.
18. The U.S. Consumer Confidence Index increases.
19. The annualized U.S. Consumer Price Index inflation rate increases.
20. The U.S. unemployment rate increases.
21. Continental Airlines' stock price increases.
22. United Parcel Service (UPS)'s stock price increases.
23. Exxon Mobil's stock price outperforms United Parcel Service (UPS)'s stock price.
24. General Motors' stock price increases.
25. IBM's stock price increases.
26. Dell's stock price outperforms Sun Microsystems' stock price.
27. Intel's stock price increases.
28. Microsoft's stock price increases by more than 10%.
29. Dell's stock price outperforms IBM's stock price.
30. Dell's stock price outperforms Apple Computer's stock price.

Appendix 2: Proof of Proposition (13)

Consider the following incoherent estimates for three absolute events.

$$(p \vee q,\ .1) \quad (\neg q,\ .1) \quad (\neg p,\ .9)$$

Let x, y, and z be the coherent replacements for $.1$, $.1$, and $.9$, respectively. Let state a assign truth to $p \vee q$ and $\neg p$, and falsity to $\neg q$. Let state b assign truth to $\neg q$ and $\neg p$, and falsity to $p \vee q$. Both of these states induce slopes of $.4$ for the original (incoherent) estimates. The situation is summarized as follows.

coherent probabilities	x	y	z	
formulas	$p \vee q$	$\neg q$	$\neg p$	
original estimates	.1	.1	.9	
state a	t	f	t	.4
state b	f	t	t	.4

For the coherent replacements x, y, z to obtain slopes as large as those in states a and b, coherent estimates must satisfy:

$$\frac{x+z}{2} - y \geq .4 \text{ and } \frac{y+z}{2} - x \geq .4, \text{ hence}$$

$$x + z - 2y \geq .8 \text{ and } y + z - 2x \geq .8, \text{ so by subtraction}$$

$$x - y - 2y + 2x \geq 0, \text{ hence } x \geq y.$$

But also $x \geq 1 - y$ by coherence, which implies that $x \geq .5$. For slope as high as achieved in b, we must thus have $\frac{y+z}{2} \geq .9$. Hence $y, z \geq .8$ and $x \geq .5$; and the latter values are incoherent. Q.E.D.

TABLE 1

Form of the events judged in the experiment

#	*form*	*example*
10	p	the U.S. retail sales rate increases
6	$p : q$	the NASDAQ Composite Index increases assuming that Microsoft's stock price increases by more than 10%
6	$p : \neg q$	the U.S. unemployment rate increases assuming that the U.S. Consumer Confidence Index does not increase
6	$p \wedge q$	U.S. 30-year fixed mortgage rates decrease and the NASDAQ Composite Index increases
6	$p \wedge \neg q$	the U.S. prime lending rate increases and the Standard & Poor's 500 Index does not increase
6	$p \vee q$	Reliant Energy's stock price increases or Exxon Mobil's stock price increases
6	$p \vee \neg q$	Wal-Mart's stock price increases or the U.S. retail sales rate does not increase
46	= *Total number of judgments*	

TABLE 2

Probabilities attributed to simple and complex events

Mean (and S.D.) for judgments types						
p	$p : q$	$p : \neg q$	$p \wedge q$	$p \wedge \neg q$	$p \vee q$	$p \vee \neg q$
.453 (.102)	.457 (.158)	.409 (.140)	.320 (.153)	.353 (.137)	.584 (.185)	.632 (.141)

Note. The mean and standard deviation are shown for the average probabilities participants assigned to events of each logical form. For example, the top left cell is the mean of the average probabilities assigned by the 47 subjects to their 10 variables. This is identical to the grand mean for all 470 estimates.

TABLE 3

Four coherence laws

(a)
$$\boxed{\begin{array}{lcl} \mathit{Prob}\,(p \wedge q) & \leq & \min\{\mathit{Prob}\,(p), \mathit{Prob}\,(q)\} \\ \mathit{Prob}\,(p \wedge q) & \geq & \mathit{Prob}\,(p) + \mathit{Prob}\,(q) - 1 \end{array}}$$

(b)
$$\boxed{\begin{array}{lcl} \mathit{Prob}\,(p \wedge \neg q) & \leq & \min\{\mathit{Prob}\,(p), 1 - \mathit{Prob}\,(q)\} \\ \mathit{Prob}\,(p \wedge \neg q) & \geq & \mathit{Prob}\,(p) - \mathit{Prob}\,(q) \end{array}}$$

(c)
$$\boxed{\begin{array}{lcl} \mathit{Prob}\,(p \vee q) & \geq & \max\{\mathit{Prob}\,(p), \mathit{Prob}\,(q)\} \\ \mathit{Prob}\,(p \vee q) & \leq & \mathit{Prob}\,(p) + \mathit{Prob}\,(q) \end{array}}$$

(d)
$$\boxed{\begin{array}{lcl} \mathit{Prob}\,(p \vee \neg q) & \geq & \max\{\mathit{Prob}\,(p), 1 - \mathit{Prob}\,(q)\} \\ \mathit{Prob}\,(p \vee \neg q) & \leq & \mathit{Prob}\,(p) - \mathit{Prob}\,(q) + 1 \end{array}}$$

Note. For example, law (a) states that the probability of a conjunction cannot exceed the probability of its conjuncts, and can not be less than one minus their sum.

TABLE 4

Average number of violations of coherence

	Tolerance		
Law	0%	5%	10%
(a)	2.96	2.49	1.94
(b)	3.47	2.87	2.06
(c)	3.13	2.51	2.09
(d)	3.57	2.79	2.21

Note. Laws (a) - (d) are stated in Table 3. In each case, the number of possible violations is 6. Tolerance of $x\%$ counts a violation only if cannot be avoided by adjusting estimates by $x\%$ or less.

TABLE 5

Average MAD for coherent approximations ($N = 47$)

application	*av. MAD*	*(S.D.)*
SAPA to all 46 judgments	.090	(.038)
LP to the 34 absolute judgments	.079	(.044)
SAPA to the 34 absolute judgments	.085	(.043)

Note. LP is linear programming. SAPA is simulated annealing applied to probability arrays.

TABLE 6

Mean quadratic penalty before and after coherent approximation ($N = 47$)

	Quadratic penalty computed for:	Mean	(S.D.)
(a)	raw estimates	.314	.077
(b)	after individual approximation by SAPA using absolute difference	.285	.076
(c)	after individual approximation by SAPA using squared deviation	.277	071
(d)	raw absolute estimates	.309	.079
(e)	absolute estimates after individual approximation by FINETTI	.272	.074
(f)	absolute estimates after individual approximation by SAPA (absolute difference)	.276	.081
(g)	after aggregate approximation by SAPA (using absolute difference)	.254	.042

Note. SAPA is simulated annealing applied to probability arrays using either absolute or quadratic distance as the proximity metric. FINETTI is de Finetti's method.

TABLE 7

Mean slope before and after coherent approximation ($N = 47$)

	Slope computed for:	Mean	$(S.D.)$
(a)	raw estimates	$.053$	$.130$
(b)	after individual approximation by SAPA	$.115$	$.129$
(c)	after aggregate approximation by SAPA	$.123$	$.092$

Note. SAPA is simulated annealing applied to probability arrays. High slopes correspond to greater accuracy.