DECISION ANALYSIS
Vol. 5, No. 4, December 2008, pp. 177–189
ISSN 1545-8490 | EISSN 1545-8504 | 08 | 0504 | 0177

informs

DOI 10.1287/deca.1080.0119

Aggregating Probabilistic Forecasts from Incoherent and Abstaining Experts

Joel B. Predd
RAND Corporation, Pittsburgh, Pennsylvania 15213, jpredd@rand.org

Daniel N. Osherson
Department of Psychology, Princeton University, Princeton, New Jersey 08544, osherson@princeton.edu

Sanjeev R. Kulkarni, H. Vincent Poor
Department of Electrical Engineering, Princeton University, Princeton, New Jersey 08544,
{kulkarni@princeton.edu, poor@princeton.edu}

Decision makers often rely on expert opinion when making forecasts under uncertainty. In doing so, they confront two methodological challenges: the elicitation problem, which requires them to extract meaningful information from experts; and the aggregation problem, which requires them to combine expert opinion by resolving disagreements. Linear averaging is a justifiably popular method for addressing aggregation, but its robust simplicity makes two requirements on elicitation. First, each expert must offer probabilistically coherent forecasts; second, each expert must respond to all our queries. In practice, human judges (even experts) may be incoherent, and may prefer to assess only the subset of events about which they are comfortable offering an opinion. In this paper, a new methodology is developed for combining expert assessment of chance. The method retains the conceptual and computational simplicity of linear averaging, but generalizes the standard approach by relaxing the requirements on expert elicitation. The method also enjoys provable performance guarantees, and in experiments with real-world forecasting data is shown to offer both computational efficiency and competitive forecasting gains as compared to rival aggregation methods. This paper is relevant to the practice of decision analysis, for it enables an elicitation methodology in which judges have freedom to choose the events they assess.

Key words: combining forecasts; probability forecasting; probability; incoherence; adjusting forecasts
History: Received on November 8, 2007. Accepted on May 28, 2008, after 1 revision. Published online in *Articles in Advance* July 31, 2008.

1. Introduction

Linear averaging is a popular method for aggregating forecasts of probability because of its simplicity, axiomatic justification, and documented empirical success; see, e.g., Genest and Zidek (1986), Clemen (1989), and Clemen and Winkler (1999) for surveys. Despite its appeal, linear averaging suffers from several limitations.

To see the issue, consider the scenario depicted in Table 1, in which three judges forecast the outcome of three events. Alice is probabilistically incoherent, because she has assigned greater probability to the event X than to the event X or Y. The unweighted average of the three judges' forecasts is similarly incoherent. More generally, it is well known that if judges forecast the same events as in Table 1, any linear-averaged aggregate is coherent if each judge is coherent. However, this is not generally true if any judge is incoherent.

Consider a second example, depicted in Table 2. The same panel convenes, but this time each judge abstains from forecasting one of the three events. Although the judges are internally coherent on their respective domains, the average of the available forecasts is incoherent in the same way that Alice was previously.

Thus, to the extent that having a coherent aggregate is desirable, linear averaging is invalid given incoherent or abstaining judges. An analyst can circumvent these problems via constrained elicitation. For example, she could construct a survey that admits forecasts for only logically independent events, in which case a

Table 1 Linear Averaging with Incoherent Judges

	Alice	Bob	Chris	Aggregate
X	**0.75**	0.50	1.00	**0.75**
Y	0.25	0.25	0.25	0.25
X or Y	**0.25**	0.75	1.00	**0.67**

judge is coherent as long as her forecasts lie between zero and one. However, this is not a general solution because it may require the analyst to ignore useful information. For example, a market analyst may predict a drop in the NASDAQ, and also assess the likelihood of a simultaneous decrease in the NASDAQ *and* in Google stock; a geopolitical expert may assess the chance of a terror attack in a *particular* city and also forecast the probability of an attack in *any* city; an atmospheric scientist could forecast an increase in carbon dioxide levels, while predicting a rise of global temperature *given* such an increase. These examples show that there is information in forecasts for logically complex events; an elicitation methodology that requires ignoring this information is suboptimal.

Alternatively, the analyst may work to instill coherence within judges through elicitation techniques; see Alpert and Raiffa (1982), von Winterfeldt and Edwards (1986), and Morgan and Henrion (1990) for discussion. Arguably, this approach is the optimal strategy, but it requires nontrivial interaction between the judge and the analyst, especially given the notoriously incoherent character of human probability judgement (Kahneman and Tversky 2000, Tentori et al. 2004). There may well be scenarios where the necessary training and communication are not feasible. Overall, it may be simpler for the analyst to be prepared to aggregate incoherent forecasts.

Abstention must also be addressed. In the example in Table 2, it would be wasteful to ignore the experts' partially specified opinions, and it would be risky to require them to complete the table with forecasts they prefer not to make.

Table 2 Linear Averaging with Abstaining Judges

	Alice	Bob	Chris	Aggregate
X	—	0.50	1.00	**0.75**
Y	0.25	—	0.25	0.25
X or Y	0.25	0.75	—	**0.50**

Table 3 An Opportunity

	Expert 1	Expert 2	Expert 3	···	Expert m
X	0.75	—	—	···	0.50
not Y	—	0.50	0.25	···	0.00
X or Y	1.00	0.75	0.70	···	—
Z	0.00	—	0.00	···	0.25
⋮	⋮	⋮	⋮	⋮	⋮
Z or W	0.75	0.00	—	···	0.50
Z if and only if X	—	0.50	0.90	···	0.10
not Z given W	0.90	—	0.50	···	0.25
V	0.50	0.75	—	···	—

In short, we believe that aggregation technology should allow for freewheeling, comfortable elicitation by accommodating incomplete and incoherent judgement. To illustrate, consider the scenario suggested by Table 3, in which m experts convene. Each assesses the chance of those events (or conditional events) about which s/he is comfortable offering an opinion. For example, Expert 1's forecast may pertain to the U.S. economy and terrorism, Expert 2's may concern terrorism and international politics, and Expert 3 may concentrate on international politics and the U.S. economy. The panel will inevitably disagree, which is to say that the union of their forecasts is not expected to be coherent (indeed, some of the judges may be internally incoherent). However, the table embodies a rich set of forecasts based on the specific competence of each judge. How should the analyst proceed to aggregate?

Osherson and Vardi (2006) propose a *coherent approximation principle* (CAP). As formalized below, CAP suggests finding a coherent forecast that is minimally different from the judges' forecasts (in a least-squares sense). CAP can be viewed as a generalization of linear averaging. In particular, if the judges are coherent and "nonabstaining," the CAP-aggregate forecast is equivalent to the unweighted-average aggregate. Unfortunately, unlike linear averaging, CAP is computationally difficult to implement in cases of interest, and in fact can be easily converted to an NP-hard decision problem.

Osherson and Vardi go on to propose a method for addressing CAP's computational challenge. Termed SAPA (Simulated Annealing over Probability Arrays), their algorithm applies to a broad class of logically complex forecasts. Although better than off-the-shelf

tools, SAPA may nonetheless take many hours to converge for modest-sized problems. Moreover, SAPA is based on an implementation of simulated annealing that requires numerous parameters to be tuned, further limiting its usability.

The purpose of this paper is to develop a more practical tool for implementing CAP, and to underscore its relevance to decision and risk analysis. Like linear averaging, our tool is quick and easy to implement (and allows judges to be "weighted" according to their credibility). And like CAP, it allows us to aggregate forecasts from incoherent and abstaining judges in a principled way. The tool thus enables a practical methodology for expert elicitation in which judges have freedom in choosing the events they assess.

1.1. Related Work

Our study extends a long tradition of inquiry into statistical methods for aggregating judgments about numerical quantities. Such methods apply a function to available estimates without a preliminary stage of consultation among judges. A standard example is the tendency of average estimates of room temperature to be closer to the true value than are most of the group's individual estimates (reported by Knight 1921, cited in Lorge et al. 1958). A comparison of statistical aggregation to methods based on discussion protocols (such as the *Delphi method*, Dalkey and Helmer 1963) is available in Hogarth (1977). See Sniezek and Henry (1989) for an analysis of the process by which groups combine individual opinions into a group judgment with associated level of confidence.

As indicated earlier, eliciting and aggregating probabilities raise special issues of intra- and interjudge consistency. They have been addressed in an extensive literature within computer science and engineering, law, philosophy, psychology, risk analysis, and statistics. A thorough survey is outside the scope of this paper. See Genest and Zidek (1986), Clemen (1989), and Clemen and Winkler (1999).

Lindley et al. (1979) offer a Bayesian approach to eliminating incoherence from a single forecast. For a Bayesian, intrajudge incoherence can be 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.

Unfortunately, the justification by Lindley et al. of the Bayesian 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). Although CAP admits a Bayesian interpretation (as discussed below), it is better motivated by the fact that it brings coherence to a body of probability estimates through minimal modification. If the modification were not minimal, we might lose whatever insight exists in the original judgments; and the panel may reject the aggregate forecast. The general idea of minimally deforming a set of beliefs to restore consistency is familiar from the copious literature on belief revision (see Gärdenfors 1988, Hansson 1999, and references cited there).

Proponents of Dempster-Shafer theory (Shafer 1976) object to probability as an idiom for belief, in part because of its inability to distinguish uncertainty from ignorance. Such a person may object to the very premise of this paper, and favor Dempster fusion rules for aggregating belief. Here we rely on abstention as an alternative expression of ignorance. Thus, CAP and the tools developed here are applicable to the setting where judges express uncertainty with probability and ignorance through abstention, and thereby afford experts more expressive freedom.

To our knowledge, the coherent approximation principle was first applied to the aggregation problem in Osherson and Vardi (2006), and SAPA is the only specialized software that addresses its computational complexity.

1.2. Outline

The remainder of this paper is organized as follows. In §2, we introduce notation. In §3, we discuss CAP and the computational challenge it presents. The algorithm we propose in response to this challenge is presented in §4. Experimental results are reported in §5,

and some extensions and implications are discussed in §6. In the appendix, we review *successive orthogonal projection algorithms* (the tools from mathematical programming on which our approach relies). For simplicity in what follows, we limit attention to probabilities of absolute events; extensions to conditional probabilities are considered at the end.

2. Preliminaries

Let Ω be a finite *outcome space*, subsets of which are called *events*. For any multiset $\mathscr{E} = \{E_1, \ldots, E_m\}$ of events, a function $f\colon \mathscr{E} \to [0,1]$ is a *forecast* that reflects how confident a judge is in the truth of events in $\mathscr{E}$.[1] For any forecast $f\colon \mathscr{E} \to [0,1]$, we let $\mathbf{f} \triangleq (f(E_1), \ldots, f(E_m)) \in [0,1]^m$, and for any event E, we let $1_E\colon \Omega \to \{0,1\}$ denote its indicator function.

Definition 1. A forecast f is (*probabilistically*) *coherent* if and only if f is consistent with some probability distribution on Ω.

The following lemma is a standard observation about coherence (de Finetti 1974; see Predd et al. 2007 for a proof).

Lemma 1. *There is a nonempty, closed, and convex set $C \subseteq [0,1]^m$ such that f is coherent if and only if $\mathbf{f} \in C$.*

Example 1. Suppose the elements of Ω are in correspondence with the 2^n realizations of a vector $(X_1, \ldots, X_n)$ of logically independent Boolean variables. In particular, $X_1, \ldots X_n$ may represent the truth or falsity of n sentences. Thus, X_1 may take value one if the sentence "Google stock outperforms the NASDAQ in the third quarter" is true, and take the value zero otherwise; similarly, X_2 may code "The average global temperature exceeds 65 degrees Fahrenheit," and X_3 may represent "A U.S. Embassy opens in Tehran." These variables can be combined in the usual way to represent disjunctive, conjunctive, negative, and other complex events.

The example evokes an outcome space composed of combinations of Boolean variables. This framework will apply to the empirical data discussed below. To simplify exposition, subsequent examples will denote complex events by their linguistic representations as in Tables 1–3. We note that our mathematical development is completely general. In particular, it applies to outcome spaces composed of combinations of arbitrary variables, not necessarily Boolean.

[1] Recall that a multiset is a setlike object that ignores order, but respects multiplicity (Weisstein 2006). For example, the multiset $\{0,0,1\} = \{0,1,0\} \neq \{0,1\}$ has three members. Intuitively, repeated members correspond to events evaluated by more than one judge. In what follows, we use the word "set" to mean "multiset" where appropriate.

Table 4 The Coherent Approximation Principle: A Running Example

$\mathscr{E}$	f_{Alice}	f_{Bob}	f_{Chris}	f_{David}
X_1	—	0.50	1.00	—
X_2	0.25	—	—	0.75
X_3	0.00	—	—	—
X_1 or X_2	1.00	0.75	—	—
X_1 and not-X_2	—	—	0.75	0.50
not X_3	—	0.75	—	—

3. The Coherent Approximation Principle

In this section, we review the coherent approximation principle for aggregating forecasts of probability. First, let us formalize a model for the panel aggregation problem.

3.1. The Panel Aggregation Problem

Suppose that on a panel of m judges, judge i provides a forecast $f_i\colon \mathscr{E}_i \to \Re$. We do not require f_i to be probabilistically coherent. Moreover, judges may abstain, which is to say that forecasts may pertain to nonidentical (but presumably related) events. In short, the forecasts are allowed to take the form suggested in Table 3. Let $\mathscr{E}$ be the (multi)set formed from pooling $\{\mathscr{E}_i\}_{i=1}^m$. *The aggregation problem requires us to construct a coherent forecast $f\colon \mathscr{E} \to [0,1]$ that reflects the opinions expressed through $\{f_i\}_{i=1}^m$.*

Example 2. Let $(X_1, \ldots, X_n)$ be as in Example 1. Consider the panel f_{Alice}, f_{Bob}, f_{Chris}, and f_{David} described by Table 4. Then, $\mathscr{E}$ is the multiset:

$$\{X_1, X_1, X_2, X_2, X_3, X_1 \text{ or } X_2, X_1 \text{ or } X_2,$$
$$X_1 \text{ and not-}X_2, X_1 \text{ and not-}X_2, \text{ not } X_3\}.$$

The aggregation problem calls for a forecast $f\colon \mathscr{E} \to [0,1]$ that best reflects the opinions of Alice, Bob, Chris, and David.

3.2. The Coherent Approximation Principle

Proposed in Osherson and Vardi (2006), the *coherent approximation principle* (CAP) suggests aggregating the panel's expertise by solving the following optimization problem:

$$\min \sum_{i=1}^{m} \sum_{E \in \mathscr{E}_i} (f(E) - f_i(E))^2 \quad \text{s.t. } f \text{ is coherent.} \tag{1}$$

Here, the optimization variable is $f\colon \mathscr{E} \to [0,1]$; the forecasts $\{f_i\}_{i=1}^m$ are the data. Notably, the output of CAP is a coherent forecast for the events in $\mathscr{E}$, but it is not necessarily a joint probability distribution over Ω.

CAP can be motivated in several ways. First, by solving (1), one finds the coherent forecast that is minimally different (with respect to squared deviation) from those provided by the panel, intuitively preserving the "information" provided by the judges while gaining probabilistic coherence. More interestingly, CAP is a strict generalization of linear averaging. In particular, when the judges are coherent and evaluate the same events (i.e., are nonabstaining), the solution to (1) is precisely the linear averaged aggregate. Next, the solution to (1) is the maximum-likelihood coherent forecast given additive white noise corrupted observations $\{f_i\}_{i=1}^m$ of a coherent source $f\colon \mathscr{E} \to [0,1]$. Finally, when there is only one incoherent forecast f_1, the solution to (1) is the so-called de Finetti point, whose Brier score (Brier 1950) is better than f_1 in all possible outcomes; see Predd et al. (2007) for a discussion of de Finetti's Theorem.[2]

3.3. The Complexity of Implementing CAP

In principle, CAP can be implemented using standard tools, because (1) is a quadratic program in $|\mathscr{E}|$ variables. To proceed, one might parameterize the feasible region with the set of probability distributions over Ω. This strategy is often computationally impractical, however, because Ω can be infeasibly large. To illustrate, consider Example 1, where $2^n - 1$ numbers are necessary to represent probability distributions over an outcome space described by a Boolean vector of length n.

Circumventing this problem would involve finding a more compact representation of the feasible region, presumably by exploiting the logical relationships between the events in $\mathscr{E}$. However, the complexity is inescapable in general. In the framework of Example 1, solving (1) is an NP-hard decision problem, because checking whether the formulas that describe events in $\{\mathscr{E}\}$ are jointly satisfiable can be reduced to checking for probabilistic coherence, which in turn can be reduced to solving (1).[3] This fact suggests that as the number of assessments $|\mathscr{E}|$ grows large and their logical complexity increases, CAP becomes intractable.

With a relaxed elicitation methodology that permits incoherent and abstaining judges, there may well be tens or hundreds of experts who provide probability forecasts for hundreds or thousands of events with a priori unforeseen logical complexity. In such circumstances, quadratic programming for CAP cannot be expected to scale up feasibly. In short, despite its conceptual simplicity, implementing CAP is computationally infeasible in cases of interest.

4. A Scalable Algorithm for Aggregation

In the previous section, we discussed how CAP is a natural extension of linear averaging, though implementing it is often computationally infeasible. In this section, we develop an efficient algorithm to address the complexity. The algorithm is quick and easy to implement like linear averaging, but like CAP, allows us to aggregate forecasts from incoherent and abstaining judges in a principled way.

Our tool exploits the idea that the logical complexity of the events assessed by human judges is usually bounded. To illustrate with Example 1, experts may be inclined to assess events expressed using no more than two Boolean variables (e.g., three term conjunctive events such as X_i or X_j or X_k may be

[2] Note that the CAP aggregate forecast would not change if a judge provided $f(\text{not } E)$ instead of $f(E)$, provided s/he obeyed the complementation principle $f(E) = 1 - f(\text{not } E)$. This invariance is due to the fact that CAP employs a quadratic distance measure. Note that $(f(\text{not } E) - p)^2 = (1 - f(E) - p)^2 = (f(E) - (1-p))^2$. So, through a linear change of variables, the optimization problem that embodies CAP would be the same regardless of which forecast was presented.

[3] See Homer and Selman (2001) for a discussion of computational complexity.

deemed inscrutable). This constraint imposes a structure that allows us to decompose (1) into a collection of subproblems, each of which can be solved quickly.

Before proceeding, recall that $\mathscr{E}$ is the multiset formed from pooling $\{\mathscr{E}_i\}_{i=1}^m$. The forecast $\hat{f}\colon \mathscr{E} \to [0,1]$ formed by pooling $\{f_i\}_{i=1}^m$ is itself a forecast. The solution to

$$\begin{aligned} \min \ & \sum_{E\in\mathscr{E}} (f(E) - \hat{f}(E))^2 \\ \text{s.t.} \ & f \text{ is coherent} \end{aligned} \tag{2}$$

is plainly equivalent to the solution to (1). That is, via pooling, CAP can be reduced to the problem of imposing coherence on a single judge. Without loss of generality, we therefore focus on (2), i.e., the case concerning a single forecast.

Example 3. Let the forecasts f_{Alice}, f_{Bob}, f_{Chris}, f_{David}, and $\mathscr{E}$ be as in Table 4. Then, $\hat{f}$ is given by:

$\mathscr{E}$	X_1	X_1	X_2	X_2	X_3	X_1 or X_2	X_1 or X_2	X_1 and not-X_2	X_1 and not-X_2	not-X_3
$\hat{f}$	0.50	1.00	0.25	0.75	0.00	1.00	0.75	0.75	0.50	0.75

Optimization (2) calls for $f\colon \mathscr{E} \to [0,1]$ that minimizes squared deviation from $\hat{f}$.

4.1. Local Coherence Constraints

Consider relaxing the requirement that a forecast $f\colon \mathscr{E} \to [0,1]$ be probabilistically coherent to the requirement that its restriction to a subset $\mathscr{F} \subseteq \mathscr{E}$ be coherent. To pursue this idea, we say that f is *locally coherent with respect to a subset* $\mathscr{F}$ if and only if f restricted to $\mathscr{F}$ is coherent (that is, if and only if $f'\colon \mathscr{F} \to [0,1]$ is coherent where $f'(E) = f(E)$ for all $E \in \mathscr{F}$). Note that "global" coherence is recovered by taking $\mathscr{F} = \mathscr{E}$, and that any coherent forecast $f\colon \mathscr{E} \to [0,1]$ must be locally coherent with respect to every subset $\mathscr{F} \subseteq \mathscr{E}$.

The following relaxation of (2) is formulated by choosing a collection of subsets $\{\mathscr{F}_\ell\}_{\ell=1}^L$.

$$\begin{aligned} \min \ & \sum_{E\in\mathscr{E}} (f(E) - \hat{f}(E))^2 \\ \text{s.t.} \ & f \text{ is locally coherent w.r.t. } \mathscr{F}_\ell \quad \forall\, \ell = 1, \ldots, L. \end{aligned} \tag{3}$$

As before, f is the optimization variable. $\hat{f}$ and $\{\mathscr{F}\}_{\ell=1}^L$ are now program data. The reason (3) relaxes (2) is that local coherence does not imply global coherence.

Linear averaging and CAP correspond to different relaxations. We illustrate with the panel exhibited by Table 5, relying on the notation of Example 1. The design corresponding to linear averaging is

$$\begin{aligned} \mathscr{F}_{\text{Avg},1} &= \{X_1, X_1\} \\ \mathscr{F}_{\text{Avg},2} &= \{X_2, X_2\} \\ \mathscr{F}_{\text{Avg},3} &= \{X_3\} \\ \mathscr{F}_{\text{Avg},4} &= \{X_1 \text{ or } X_2, X_1 \text{ or } X_2\}. \\ \mathscr{F}_{\text{Avg},5} &= \{X_1 \text{ and not-}X_2, X_1 \text{ and not-}X_2\}. \\ \mathscr{F}_{\text{Avg},6} &= \{\text{not } X_3\}. \end{aligned}$$

Under this design, the relaxation incorporates few coherence constraints, and thus the solution to (3) may poorly approximate the CAP-aggregate. CAP requires global coherence and corresponds to the design

$$\begin{aligned} \mathscr{F}_{\text{CAP}} = \{&X_1, X_1, X_2, X_2, X_3, X_1 \text{ or } X_2, X_1 \text{ or } X_2, \\ &X_1 \text{ and not-}X_2, X_1 \text{ and not-}X_2, \text{ not-}X_3\}. \end{aligned}$$

As we will soon see, alternative relaxations can be motivated by the desire to trade-off the efficiency of linear averaging against the goal of finding a minimally modified coherent aggregate.

4.2. A Scalable Aggregation Algorithm

Note that the solution to the relaxation (3) permits a geometric interpretation as a projection onto the intersection of L sets. In particular, (3) can be rewritten as

$$\begin{aligned} &\min (\|\mathbf{f} - \hat{\mathbf{f}}\|_2)^2 \\ &\text{s.t.} \quad \mathbf{f} \in \bigcap_{\ell=1}^{L} C_\ell, \end{aligned}$$

where $C_\ell = \{\mathbf{f}\colon f \text{ is locally coherent with respect to } \mathscr{F}_\ell\}$ and $\|\cdot\|_2$ denotes the Euclidean norm. Using Lemma 1, it is easy to see that C_ℓ is nonempty, closed, and convex. Thus, well-studied successive orthogonal projection (SOP) algorithms can be applied to iteratively approximate the solution to (3). This observation forms the basis for our aggregation tool. We refer the reader to the appendix for a brief introduction to an SOP algorithm.

Table 5 SAA: A Scalable Algorithm for Aggregation

Input:	Forecast $\hat{f}: \mathscr{E} \to [0, 1]$, T
Initalize:	$f_0 = \hat{f}$.
Step 1:	Design $\{\mathscr{F}_\ell\}_{\ell=1}^L$ with $\mathscr{F}_\ell \subseteq \mathscr{E}$ for $\ell = 1 \ldots, L$.
Step 2:	for $t = 1, \ldots, T$
	$\ell_t = t \bmod L$
	$f_t := \arg\min \sum_{E \in \mathscr{E}} (f(E) - f_{t-1}(E))^2$
	s.t. $f: \mathscr{E} \to [0, 1]$ is locally coherent w.r.t. $\mathscr{F}_{\ell_t}$.
Output:	f_T.

Our algorithm, termed *SAA* in what follows, is detailed in Table 5. Note that in *SAA* f_t: $\mathscr{E} \to [0, 1]$ is uniquely specified by:

$$\begin{aligned} f_t(E) &= f_{t-1}(E) \quad \text{for all } E \notin \mathscr{F}_{\ell_t} \\ f_t(E) &= f^\star(E) \quad \text{for all } E \in \mathscr{F}_{\ell_t} \\ f^\star &= \arg\min \sum_{E \in \mathscr{F}_{\ell_t}} (f(E) - f_{t-1}(E))^2 \\ &\text{s.t. } f: \mathscr{F}_{\ell_t} \to [0, 1] \text{ is coherent.} \end{aligned} \tag{4}$$

Hence, the computation in the inner loop amounts to solving (4), an optimization over forecasts f: $\mathscr{F}_{\ell_t} \to [0, 1]$.

Crucial to *SAA* is Step 1, designing $\{\mathscr{F}_\ell\}_{\ell=1}^L$. Intuitively, the fewer events that each $\mathscr{F}_\ell$ contains, the faster the inner computation runs; standard optimization techniques may be applied to efficiently solve (4) if C_ℓ admits a compact parameterization. On the other hand, as $\mathscr{F}_\ell$ get larger, a richer set of coherence constraints are represented, and thus the solution to (3) more closely approximates the CAP-aggregate (2). To illustrate, the highly local approach corresponding to linear averaging can be implemented very quickly, essentially because $C_{\text{Avg},\,\ell}$ can be parameterized with only one variable for $\ell = 1, \ldots, 6$. In contrast, as discussed above, the CAP design may well be computationally infeasible. Thus, when designing $\{\mathscr{F}_\ell\}_{\ell=1}^L$, one must strike a balance between *coherence* and *speed*.

4.3. Designing Local Coherence Constraints to Trade Speed and Coherence

The preceding discussion places linear averaging and CAP at opposite extremes of a speed-coherence trade-off. A compromise may be struck through a more nuanced design of local coherence constraints. Alhough the choice of a design is ultimately a heuristic decision, in practical settings it is plausible to rely on the way events are described to the judge.

To illustrate with the panel exhibited in Table 5, consider the following design, where conjunctive events are grouped with their conjuncts, disjunctive events are grouped with their disjuncts, and negative events are grouped with their complements.

$$\mathscr{F}_{\text{or}} = \{X_1, X_1, X_2, X_2, X_1 \text{ or } X_2, X_1 \text{ or } X_2, X_1, X_1\}$$

$$\mathscr{F}_{\text{and}} = \{X_1, X_1, X_2, X_2, X_1 \text{ and not-}X_2, X_1 \text{ and not-}X_2\}.$$

$$\mathscr{F}_{\text{not}} = \{X_3, \text{ not-}X_3\}.$$

This design represents a stricter set of coherence constraints than linear averaging, because

$$\mathscr{F}_{\text{Avg},1}, \mathscr{F}_{\text{Avg},2}, \mathscr{F}_{\text{Avg},4} \subseteq \mathscr{F}_{\text{or}}$$

$$\mathscr{F}_{\text{Avg},5} \subseteq \mathscr{F}_{\text{and}}$$

$$\mathscr{F}_{\text{Avg},6}, \mathscr{F}_{\text{Avg},3} \subseteq \mathscr{F}_{\text{not}}.$$

However, $\{\mathscr{F}_{\text{or}}, \mathscr{F}_{\text{and}}, \mathscr{F}_{\text{not}}\}$ imposes fewer coherence constraints than the design corresponding to CAP. Because each subset pertains to at most two Boolean variables, C_{or}, C_{and}, and C_{not} each may be compactly parameterized by at most three free variables, which is fewer than needed for the CAP design.[4] With this design, *SAA* promises to retain the computational simplicity of linear averaging.

One can envision larger scenarios, pertaining to hundreds of Boolean variables and thousands of events. In such a setting, aggregating using linear averaging may go too far in sacrificing coherence, and the CAP approach will be intractable. If, as in the preceding examples, each of the underlying events pertains to no more than two Boolean variables, implementing *SAA* with a scalable design might result in an acceptable level of coherence attainable in a practical amount of time. These ideas are validated empirically in the experiments reported below.

4.4. A Performance Guarantee

Suppose that after learning the true outcome $\omega_t \in \Omega$, the accuracy of a forecast f is assessed using the *Brier*

[4] For example, C_{or} may be parameterized by representing joint distributions over (X_1, X_2), which requires three numbers.

score (Brier 1950) or "quadratic penalty," defined as follows:

$$\mathrm{QP}(f, \omega_t) = \sum_{E \in \mathscr{E}:\, \omega_t \in E} (1 - f(E))^2 + \sum_{E \in \mathscr{E}:\, \omega_t \notin E} f(E)^2. \quad (5)$$

SAA guarantees stepwise improvement in accuracy (lower penalty) as measured by the Brier score, and converges as $T \to \infty$ to a forecast that is locally coherent with respect to all designed subsets. Theorem 1 formalizes this fact.

THEOREM 1. *Let* f_T *be as defined by* SAA (*Table* 5). *Then*,

(1) $QP(f_T, \omega) \leq QP(f_{T-1}, \omega)$ *for all* $\omega \in \Omega$ *and all* $T \geq 1$,

(2) f_T *converges* (*i.e.*, $\mathbf{f}_T$ *converges in norm*),

(3) *and* $\lim_{T\to\infty} f_T$ *is locally coherent w.r.t.* $\mathscr{F}_l$ *for* $l = 1, \ldots, L$.

That is, forecast accuracy is preserved or improved at each step, as *SAA* converges to local coherence. The proof of Theorem 1 follows from standard analysis of SOP algorithms (see Theorem 2 in the appendix) after noticing that for any $\omega \in \Omega$, $\mathbf{f}_\omega \triangleq (1_{E_1}(\omega), \ldots, 1_{E_m}(\omega)) \in \bigcap_{\ell=1}^{L} C_\ell$ and that $\mathrm{QP}(f_T, \omega) = (\|\mathbf{f}_T - \mathbf{f}_\omega\|_2)^2$.

4.5. Comments

Depending on the design of $\{\mathscr{F}_l\}_{l=1}^{L}$, the output of *SAA* need not be coherent, because it approximates (3), a relaxation of CAP. However, for any design $\{\mathscr{F}_l\}_{l=1}^{L}$ the output will be closer than the input to coherence, because it will satisfy a set of local coherence constraints. This fact is formalized by Theorem 1.

Note that the Brier score of the output of *SAA* can only be improved (lowered) by adding local coherence constraints. This fact follows from the same argument[5] used to prove Theorem 1, and further motivates our approach.

The present proposal decomposes CAP into a sequence of small problems, which can be quickly and independently solved with out-of-the-box software, or more specialized tools like SAPA. Of course, the number of iterations T is a design parameter that must be tuned. Tuning is facilitated by Theorem 1, however, because performance is monotonic in T.

[5] Apply Theorem 1 to the case when *SAA* employs just two local coherence constraints, $\mathscr{F}$ and $\mathscr{F} \cup \mathscr{G}$.

Note that in Step 2, the local coherence constraints are addressed in sequence. The precise ordering is unimportant and parallelism may be introduced. In particular, two projections can occur simultaneously as long as they pertain to disjoint sets of events. More generally, *SAA* can be extended to a range of nonsequential, even random, orderings. See Censor and Zenios (1997) for related extensions of the SOP algorithm.

Despite the strong performance guarantees described by Theorem 1, *SAA* only approximates the solution to (3). To solve (3) exactly, one may employ Dykstra's method (Dykstra 1983), an iterative alternative to the SOP algorithm. When applied in the present context of aggregation, the algorithm will converge as $T \to \infty$ to the solution to (3), but it will do so without the stepwise improvement in accuracy afforded by *SAA*.

5. Experiments

In the previous section, we developed and analyzed *SAA*, an algorithm for (approximately) implementing CAP. In this section, we empirically validate the implementation, focusing on three main issues: (i) the computational efficiency of the algorithm in practice, (ii) how well the algorithm approximates the input forecasts, and (iii) the algorithm's effect on forecasting accuracy.

5.1. The Data

Five previously collected data sets were used in these experiments. The STCK database was first published in Osherson and Vardi (2006) and contains forecasts made by MBA students at Rice University on events pertaining to 10 stocks in the third quarter of 2000; the FIN database is documented in Batsell et al. (2002) and summarizes forecasts made by students at Rice on events related to various economic indicators in the fourth quarter of 2001; the NBA1 and NBA2 data sets appeared in Batsell et al. (2002) and detail forecasts made by self-identified basketball enthusiasts regarding the outcome of two Houston Rockets National Basketball Association games; the HSTN data set Hendrix et al. (2005) contains forecasts made by Houston homeowners on events pertaining to the local real-estate market and pollution.

Table 6 Summary of Data in Aggregation Experiments

	STCK	FIN	NBA1	NBA2	HSTN
Subjects	47	31	29	36	17
Basic events	30	10	10	10	10
Judgments/Agg.	1,598	1,054	986	1,224	578
Events/Agg.	1,000	308	405	341	259

In each of the five data sets, subjects were asked to assess the probability of 34 randomly selected basic (10) and complex (24) events. Each basic event was a Boolean variable, e.g., "the U.S. Consumer Confidence Index increases in the third quarter of 2000." Complex events were constructed from the basic events allocated to a given subject. They had one of the following forms: X and Y, X and not-Y, X or Y, X or not-Y. Each form appeared equally often in a given subjects' collection of 24 complex events.

The number of subjects (i.e., the size of the panel) per data set is summarized in Table 6, along with the total number of basic events from which the forecasted events were constructed. In Table 6, "Judgments/Agg" describes the total number of judgments made by the panel (i.e., $|\mathscr{E}|$), and "Events/Agg" indicates the number of *unique* events assessed. Note that Events/Agg is less than Judgments/Agg, because each event is typically assessed by multiple judges.

Observe that subjects in these experiments "abstained" in the involuntary sense of not being asked every question in the study. The random assignment of events to judges nonetheless had the effect of creating a set of sparse forecasts similar to the one suggested by Table 3. These data sets are therefore suitable for experimenting with SAA.

5.2. The Method

To each of the data sets, we applied *SAA*, the aggregation algorithm detailed in §4. In all cases, the design of local coherence constraints was based on the complex events figuring in the experiment. Specifically, for each complex event E, the design includes the set of all copies of E, along with all copies of its constituent basic events. For example, corresponding to the event X or not-Y, one component of the design is the set consisting of all copies of X, Y, and X or not-Y. This design is an obvious generalization of $\{\mathscr{F}_{\text{or}}, \mathscr{F}_{\text{and}}, \mathscr{F}_{\text{not}}\}$, illustrated in the previous section, wherein conjunctions are grouped with their conjuncts and disjunctions are grouped with their disjuncts. In what follows, we refer to our design as "scalable."

For every forecast reported in each database, the truth values of the corresponding events are known. This allows us to assess the accuracy of various forecasts a posteriori. Accuracy is measured using the *Brier score* and *slope* (Yates 1990). The Brier score is defined in Equation (5). The slope of a forecast of $f\colon \mathscr{E} \to [0, 1]$ is defined as

$$\frac{1}{m_T} \sum_{E\in\mathscr{E}:\, E \text{ is TRUE}} f(E) - \frac{1}{m - m_T} \sum_{E\in\mathscr{E}:\, E \text{ is FALSE}} f(E),$$

where m_T denotes the number of true events in $\mathscr{E}$. Slope functions as a reward, inasmuch as higher slopes indicate more accurate forecasts (in contrast to the Brier score, which acts as a penalty).

We assess the accuracy of forecasts in four contexts.

- *Raw.* The average accuracy of the judges' unprocessed forecasts is measured.
- *Individual.* After eliminating intrajudge incoherence (i.e., after running *SAA* on each individual judge under the scalable design), the average accuracy of the judges' forecasts is reported.
- *Aggregate.* After eliminating inter- and intrajudge incoherence (i.e., after running *SAA* on the pool of all judgments under the scalable design), the average accuracy of the judges' forecasts is measured.
- *Linear Average*: After replacing each forecast for a given event with the group's average for that event, the average accuracy of the judges' forecasts is assessed.

Note that slope reported for *linear averaging* is the same as slope reported for *raw*.

5.3. Computational Efficiency

Figures 1 and 2 detail the average Brier score achieved by the panel versus the number of iterations (T) made by *SAA* in the *Individual* and *Aggregate* cases, respectively. The monotonicity of these plots is predicted by Theorem 1. In both cases and in every data set, *SAA* converges within 10 iterations through the forecasts.

From a computational perspective, the most interesting data set is the STCK database, because it contains the largest number of unique events per aggregate, the most basic events, and the greatest

Figure 1 **Individual: Average Brier Score vs. T**

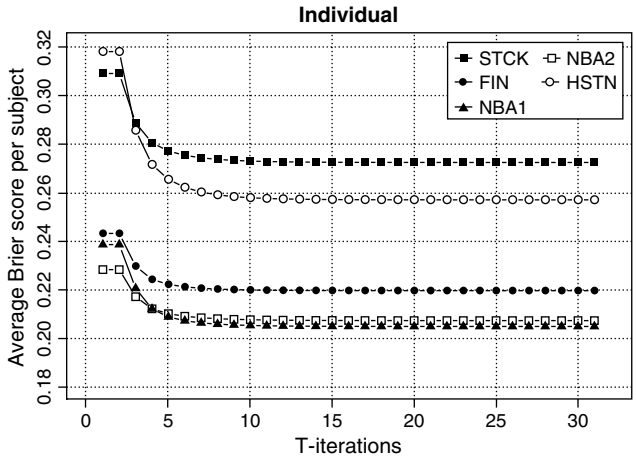
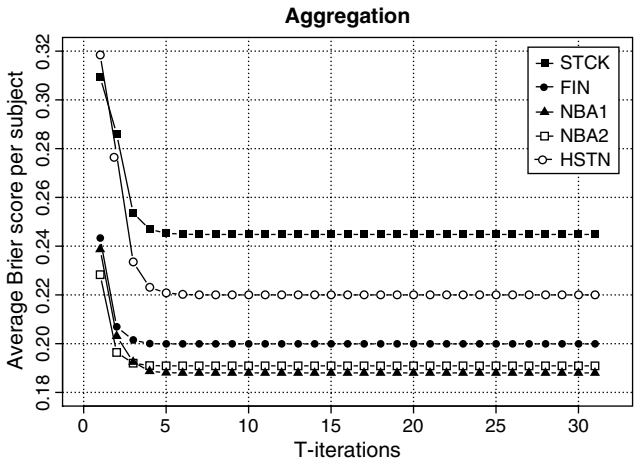


Figure 2 **Aggregate: Average Brier Score vs. T**



number of subjects (see Table 7). On a 1 GHz PowerPC G4, *SAA* required approximately 10.0 s to converge when applied to the 47 judges' pooled forecast (i.e., to compute *Aggregate*). In contrast, the rival method SAPA (Osherson and Vardi 2006) required several hours. The time required to eliminate incoherence from individual judges was less than 0.6 s (i.e., to compute *Individual*).

5.4. Approximation of Input Forecasts

For the STCK database, Osherson and Vardi (2006) report that SAPA competes with CAP in approximating input forecasts: The mean absolute deviation[6] between subjects' original forecasts and the SAPA-processed forecasts (*Individual*) was 0.085 in comparison to "optimal" CAP, which was 0.079. On the same data set, *SAA* records 0.064 MAD. (Because *SAA* is designed to solve a relaxation of CAP, we expect *SAA* to more closely approximate the input forecasts than both SAPA and CAP.)

5.5. Forecasting Accuracy

Osherson and Vardi (2006) report three empirical findings. First, they observe that eliminating intrajudge incoherence improves the forecasting accuracy of individual judges (i.e., *Individual* is better than *Raw*). Second, they observe that panel aggregation improves the forecasting accuracy of panel members (i.e., *Aggregate* improves *Raw*). Finally, Osherson and Vardi (2006) report that aggregation improves the accuracy of panel members as compared to incoherence-corrected forecasts (i.e., *Aggregate* improves *Indivdual*). These findings are anticipated by Theorem 1 when accuracy is assessed using the Brier score. However, the empirical findings were upheld under alternative accuracy measurements including slope. The present experiment examines whether Osherson and Vardi's observations hold up when using *SAA*.

Tables 7 and 8 summarize the result for Brier score and slope, respectively. We see that the findings of Osherson and Vardi are retained when using *SAA*, except that the aggregate slopes are not consistently higher than for the individual application of our algorithm. The aggregate slope was higher than individual slope in the largest of our data sets (namely, STCK, involving 30 variables); it was lower for the remain-

Table 7 **Forecasting Accuracy: Brier Score**

	STCK	FIN	NBA1	NBA2	HSTN
Raw	0.309	0.243	0.239	0.228	0.318
Individual	0.273	0.220	0.205	0.207	0.257
Aggregate	0.245	0.200	0.188	0.191	0.220
Linear avg.	0.286	0.207	0.203	0.196	0.234

Table 8 **Forecasting Accuracy: Slope**

	STCK	FIN	NBA1	NBA2	HSTN
Raw	0.064	0.153	0.140	0.141	0.129
Individual	0.109	0.172	0.186	0.169	0.210
Aggregate	0.114	0.153	0.173	0.150	0.202

[6] The mean absolute deviation (MAD) between forecasts $f\colon \mathscr{E} \to [0,1]$ and $g\colon \mathscr{E} \to [0,1]$ is $(1/|\mathscr{E}|)\sum_{E\in\mathscr{E}} |f(E) - g(E)|$.

ing, smaller data sets (each involving just 10 variables). Furthermore, SAPA and *SAA* yield comparable forecasting gains in absolute terms. To illustrate with the STCK data set, SAPA yields an average Brier score of 0.276 (*Indivdual*), whereas the CAP calculation computed using quadratic programming yields 0.272. The value for *SAA* is 0.273, slightly superior to SAPA.

We see that *SAA* provides a significant computational speed-up, while achieving forecasting gains similar to CAP.

6. Conclusion

In summary, linear averaging is of limited use for judgement aggregation in the context of incoherent or abstaining judges. The coherent approximation principle (CAP) has wider applicability, yet suffers from computational intractability in cases of interest. We have proposed a unified framework that positions CAP and linear averaging at opposite extremes of a speed-coherence trade-off, and suggests a principled methodology for compromise. By exploiting the logical simplicity of events typically assessed by human judges, our aggregation tool (*SAA*) offers an acceptable level of coherence in practical amounts of time. Moreover, *SAA* enjoys provable performance guarantees. Empirically, several experiments document the computational efficiency of *SAA* along with forecasting gains similar to CAP.

In cases where probabilistically coherent forecasts are necessary, *SAA* can be used as a stepping stone to deriving a fully coherent forecast. For example, one possibility is to incrementally add local coherence constraints until *SAA* arrives at a coherent forecast.[7] The performance of such schemes will depend on the structure of the events in question. However, adding additional constraints will improve the Brier score (as established by Theorem 1) and in general increase computation time.

CAP has been formulated using a quadratic distance measure. It can be naturally generalized to *Bregman divergences*, a class of distance measures that include (weighted) Euclidean distance and binary relative entropy as special cases (Bregman 1967). *SAA* can be generalized to an arbitrary Bregman divergence by applying Bregman's algorithm (Censor and Zenios 1997), which generalizes the SOP algorithm to Bregman divergences. Thus, the methodology developed here has broad applicability.

Forecasts of conditional probability are often easily elicited from human experts (Pearl 1988). Because conditional probabilities are specified as ratios, the corresponding optimization problem underlying CAP is nonconvex and therefore may have local minima. Two responses to this difficulty may be envisioned. CAP may be retained along with the computational burden engendered by conditional probabilities. Alternatively, CAP can be modified to ensure convexity even with conditional probabilities, but with loss of simplicity in the conception of "proximity" to the panel's judgments. The former approach requires algorithmic development, whereas the latter calls for axiomatic justification for alternatives to CAP. We expect *SAA* to be a useful component in both enterprises.

When considering expert opinion, decision makers may wish to incorporate information about the credibility of individual judges. For this purpose, the weights in a linear average can be adjusted to favor the forecasts of trusted experts. Both CAP and SAA can be similarly modified to weight certain forecasts more than others. In addition, by providing judges with the flexibility to abstain, CAP (and therefore SAA) opens the door to self-evaluation of credibility, because a judge may simply decline to estimate the probability of events beyond her expertise.

In practice, large problems sometimes decompose in a way that allows them to be addressed piecemeal. As we argued in the introduction, however, this is not universally the case, and experts may be called upon to estimate large sets of logically interrelated events. In such circumstances, incoherence and abstention are expected, and CAP is infeasible. Our proposed method (SAA) is designed to overcome these obstacles. Without a method like SAA, practitioners will be tempted to avoid eliciting expert assessments of large numbers of logically interrelated events in order to mold their data into a form suitable for linear averaging. Tools like SAA open the door to less-constrained elicitation procedures by handling the attendant complexity of aggregation. It is an open question whether panels of experts who have freedom to choose the

[7] We thank a reviewer for suggesting this point.

Table A.1 The SOP Algorithm

Initialize:	$\mathbf{x}_0 := \hat{\mathbf{x}}$
Iterate:	$\mathbf{x}_{n+1} := P_{C_{(n \bmod L)+1}}(\mathbf{x}_n)$

events they assess provide more informative judgments. In any case, SAA offers an efficient means of addressing the question.

In summary, we have argued that aggregation technology should allow for freewheeling, comfortable elicitation by accommodating incomplete and incoherent judgement. *SAA* makes this policy feasible, giving judges the freedom to choose the events they assess.

Appendix. Successive Orthogonal Projection Algorithms

Successive orthogonal projection algorithms are a well-studied tool from mathematical programming that constitute a key part of our approach. To illustrate, let $C_1, \ldots, C_L$ be closed and convex subsets of $\Re^m$, whose intersection $C = \bigcap_{\ell=1}^{L} C_\ell$ is nonempty. For any $\hat{\mathbf{x}} \in \Re^m$, let $P_{C_\ell}(\hat{\mathbf{x}})$ denote the Euclidean least-squares projection of $\hat{\mathbf{x}}$ onto C_ℓ, i.e.,

$$P_{C_\ell}(\hat{\mathbf{x}}) := \arg\min_{\mathbf{x} \in C_\ell} (\|\mathbf{x} - \hat{\mathbf{x}}\|_2)^2.$$

Here, $\|\cdot\|_2$ denotes the Euclidean norm. Depicted in Table A.1, the (unrelaxed) successive orthogonal projection (SOP) algorithm (Censor and Zenios 1997) provides a way to approximate $P_C(\cdot)$ given $\{P_{C_\ell}(\cdot)\}_{\ell=1}^{L}$. In words, the SOP algorithm successively and iteratively projects onto each of the subsets. Much of the behavior of this algorithm can be understood through Theorem 2, the proof of which can be found in Censor and Zenios (1997).

THEOREM 2. *Let $C, C_1, \ldots, C_L$ be as above, and let $\mathbf{x}_n$ be defined as in the SOP algorithm. Then*

(1) $\|\mathbf{x}_n - \mathbf{x}\|_2 \le \|\mathbf{x}_{n-1} - \mathbf{x}\|_2$ *for all* $\mathbf{x} \in C$ *and all* $n \ge 1$,
(2) $\mathbf{x}_n$ *converges in norm*,
(3) *and* $\lim_{n\to\infty} \mathbf{x}_n \in C$.

Note that for a given design $\{\mathcal{F}_\ell\}_{\ell=1}^{L}$, SAA (Table 5) is exactly the SOP algorithm (Table A.1) written in the formalism of the panel aggregation problem.

Acknowledgments

This research was completed while the first author was a Ph.D. candidate at Princeton University. This research was supported in part by the Army Research Office under Grant DAAD19-00-1-0466, in part by the U.S. Army Pantheon Project, in part by Draper Laboratory under Grant IR&D 6002, and in part by the National Science Foundation under Grants CCR-0020524, CCR-0312413, and IIS-9978135. The second author acknowledges the Luce Foundation. The authors thank two anonymous referees for their comments.

References

Alpert, M., H. Raiffa. 1982. A progress report on the training of probability assessors. D. Kahneman, P. Slovic, A. Tversky, eds. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, New York, 294–305.

Batsell, R., L. Brenner, D. Osherson, S. Tsavachidis, M. Y. Vardi. 2002. Eliminating incoherence from subjective estimates of chance. *Proc. 8th Internat. Conf. Principles of Knowledge Representation and Reasoning* (*KR 2002*), Toulouse, France.

Bregman, L. M. 1967. The relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *U.S.S.R. Comput. Math. Math. Phy.* **78**(384), 200–217.

Brier, G. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Rev.* **78** 1–3.

Censor, Y., S. A. Zenios. 1997. *Parallel Optimization: Theory, Algorithms, and Applications*. Oxford University Press, New York.

Clemen, R. T., R. Winkler. 1999. Combining probability distributions from experts in risk analysis. *Risk Anal.* **19** 187–203.

Clemen, R. T. 1989. Combining forecasts: A review and annotated bibliography. *Internat. J. Forecasting* **5** 559–583.

Dalkey, N. C., O. Helmer. 1963. An experimental application of the delphi method to the use of experts. *Management Sci.* **9** 458–467.

de Finetti, B. 1974. *Theory of Probability*, Vol. 1. John Wiley and Sons, New York.

Dykstra, R. L. 1983. An algorithm for restricted least squares regression. *J. Amer. Statist. Assoc.* **78**(384) 837–842.

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

Genest, C., J. Zidek. 1986. Combining probability distributions: A critique and an annotated bibliography. *Statist. Sci.* **1**(1) 114–135.

Hansson, S. O. 1999. *A Textbook of Belief Dynamics: Theory Change and Database Updating*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Hendrix, M. E., P. R. Hartley, D. Osherson. 2005. Real estate values and air pollution: Measured levels and subjective expectations. Working paper, Rice University, Houston, TX.

Hogarth, R. M. 1977. Decision making and change in human affairs. *Methods for Aggregating Opinions*. D. Reidel, Dordrecht, Holland, 231–256.

Homer, S., A. L. Selman. 2001. *Computability and Complexity Theory*. Springer, New York.

Kahneman, D., A. Tversky, eds. 2000. *Choices, Values, and Frames*. Cambridge University Press, New York.

Lindley, D. V., A. Tversky, R. V. Brown. 1979. On the reconciliation of probability assessments. *J. Royal Statist. Soc. A* **142**(Part 2) 146–180.

Lorge, I., D. Fox, J. Davitz, M. Brenner. 1958. A survey of studies contrasting the quality of group performance and individual performance. *Psych. Bull.* **55** 337–372.

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

Osherson, D., M. Y. Vardi. 2006. Aggregating disparate estimates of chance. *Games Econom. Behav.* **56**(1) 148–173.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco.

Predd, J., R. Seiringer, E. H. Lieb, D. Osherson, V. Poor, S. Kulkarni. 2007. Probabilistic coherence and proper scoring rules. Available at: http://www.citebase.org/abstract?id=oai:arXiv.org:0710.3183.

Shafer, G. 1976. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ.

Sniezek, J. A., R. A. Henry. 1989. Accuracy and confidence in group judgment. *Organ. Behav. Human Decision Processes* **43** 1–28.

Tentori, K., N. Bonini, D. Osherson. 2004. The conjunction fallacy: A misunderstanding about conjunction? *Cognitive Sci.* **28**(3) 467–477.

von Winterfeldt, D., W. Edwards. 1986. *Decision Analysis and Behavioral Research*. Cambridge University Press, New York.

Weisstein, E. W. 2006. Multiset. From MathWorld–A Wolfram Web Resource. Retrieved October 10, 2007, http://mathworld.wolfram.com/Multiset.html.

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