Decision Analysis

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Probabilistic Coherence Weighting for Optimizing Expert Forecasts

Christopher W. Karvetski, Kenneth C. Olson, David R. Mandel, Charles R. Twardy

To cite this article:
Christopher W. Karvetski, Kenneth C. Olson, David R. Mandel, Charles R. Twardy (2013) Probabilistic Coherence Weighting for Optimizing Expert Forecasts. Decision Analysis 10(4):305-326. http://dx.doi.org/10.1287/deca.2013.0279

DECISION ANALYSIS
Vol. 10, No. 4, December 2013, pp. 305–326
ISSN 1545-8490 (print) | ISSN 1545-8504 (online)

http://dx.doi.org/10.1287/deca.2013.0279

Probabilistic Coherence Weighting for Optimizing Expert Forecasts

Christopher W. Karvetski, Kenneth C. Olson
Department of Applied Information Technology, George Mason University, Fairfax, Virginia 22030
{ckarvetski@gmail.com, kolson8@gmu.edu}

David R. Mandel
Socio-Cognitive Systems Section, DRDC Toronto; and Department of Psychology, York University,
Toronto, Ontario M3J 1P3, Canada, drmandel66@gmail.com

Charles R. Twardy
Command, Control, Communications, Computing, and Intelligence Center, George Mason University,
Fairfax, Virginia 22030, ctwardy@gmu.edu

Methods for eliciting and aggregating expert judgment are necessary when decision-relevant data are scarce. Such methods have been used for aggregating the judgments of a large, heterogeneous group of forecasters, as well as the multiple judgments produced from an individual forecaster. This paper addresses how multiple related individual forecasts can be used to improve aggregation of probabilities for a binary event across a set of forecasters. We extend previous efforts that use probabilistic incoherence of an individual forecaster's subjective probability judgments to weight and aggregate the judgments of multiple forecasters for the goal of increasing the accuracy of forecasts. With data from two studies, we describe an approach for eliciting extra probability judgments to (i) adjust the judgments of each individual forecaster, and (ii) assign weights to the judgments to aggregate over the entire set of forecasters. We show improvement of up to 30% over the established benchmark of a simple equal-weighted averaging of forecasts. We also describe how this method can be used to remedy the "fifty–fifty blip" that occurs when forecasters use the probability value of 0.5 to represent epistemic uncertainty.

Key words: probabilistic coherence; forecast aggregation; crowdsourcing; linear opinion pool; fifty–fifty blip; practice
History: Received on February 21, 2013. Accepted by Rakesh Sarin on July 26, 2013, after 1 revision.

1. Introduction

1.1. Aggregating Forecaster Judgment

Decision makers often rely on the subjective judgment and expertise of forecasters when little to no decision-relevant "hard" data exist. Methods for eliciting and aggregating the judgments of multiple forecasters have proven to be valuable tools for improving the accuracy of the judgments in a variety of engineering and other settings (Cooke and Goossens 2008). Clemen and Winkler (1999) give an overview describing the combination of expert judgment in the context of risk analysis.

When forecasting the occurrence of a future event, such as the outcome of an upcoming election, each forecaster provides a subjective probability distribution concerning the resolution of the event. For binary events, the distribution is typically a single probability value, and the resolution value is *one* if the event occurs or *zero* if the event does not occur.

Although more sophisticated Bayesian and classical methods of aggregation have been proposed (e.g., Merrick 2008, Cooke and Goossens 2008), a simple, equal-weighted averaging of probability distributions across the crowd of forecasters, known as a linear opinion pool (LINOP), is generally considered the benchmark for aggregation (Clemen 2008, Clemen and Winkler 1999). Consider, for example, the Aggregative Contingent Estimation (ACE) Program, which began in 2010 as a multiyear, six-team forecasting challenge sponsored by the Intelligence Advanced Research Projects Activity (IARPA). The goal of the ACE Program is, through the development of advanced forecasting techniques, to dramatically enhance the accuracy of forecasts for a broad range of event types, compared to a linear opinion pool.

Nevertheless, it is reasonable to believe that in a large crowd, particular forecasters will produce more accurate judgments than other forecasters, and giving more weight to these forecasters will improve upon the LINOP (Genest and McConway 1990). A key challenge is identifying these better forecasters before resolution. Previous weighting schemes have used forecaster performance data on resolved questions or *seed questions* to generate forecaster weights (Cooke and Goossens 2008, Cooke 1991). However, these approaches require the existence of past performance data and they assume that the data will be a good indicator of future forecasting performance.

Recent studies have shown that weighting forecasters by the degree to which their forecasts are probabilistically coherent can improve upon a LINOP (Tsai and Kirlik 2012, Wang et al. 2011). Probabilistic coherence implies that the distribution over the events in a probability space do not violate the basic axioms of probability (De Finetti 1990, Kolmogorov 1956). Importantly, probabilistic coherence is the only evaluation method for an individual's probabilities that can be done before resolution (Lindley et al. 1979).

Although coherence weighting within aggregation has been shown to improve over a LINOP, a facile approach has never been developed and tested for simple events. That is, given a binary event A, it would be useful to know (a) what the best judgments are to elicit in addition to $P(A)$ to obtain a measure of probabilistic coherence, (b) if the related judgments should be elicited independently or concurrently, and (c) how the degree of coherence should be converted to an aggregation weight.

1.2. Overview of Paper

In this paper, using the data from two studies, we build on past efforts and develop and test an approach for eliciting small sets of related probabilities from individual forecasters, which are used to aggregate the probabilities over the crowd of forecasters. In particular, by giving greater weight to more coherent forecasters in a weighted averaging, we expect to significantly improve upon a LINOP using a relatively facile and highly feasible approach. Within the approach, we describe the importance of generating independent or "spaced" (as opposed to concurrent) intraforecaster judgments, as our method is designed to capitalize on extra incoherence produced by forecasters making independent judgments.

The rest of the paper is organized as follows: In the next section, we describe related literature for coherence weighting. We review key principles of forecast elicitation and aggregation in this second section as well. In §3, we outline our experimental methods, and §4 describes the results and insights from the two studies. In §5, we discuss the results, offer conclusions, and suggest future research.

2. Related Literature

2.1. Coherentization and Coherence Weighting

Methods of elicitation and aggregation of judgment, including the LINOP, are usually applied to a set of at least two forecasters, and are sometimes applied to a large "crowd" of forecasters (i.e., "crowdsourcing"; Surowiecki 2005). These methods, however, have also been applied to an individual forecaster. In this latter case, the individual forecaster is prompted to produce multiple, independent estimates for the same unknown parameter, and these estimates are combined to yield, on average, a more accurate judgment.

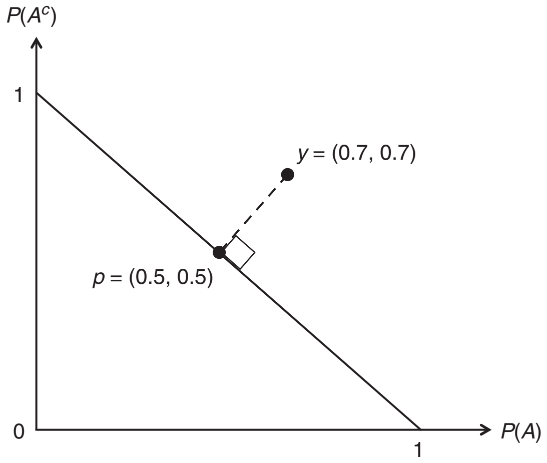
Herzog and Hertwig (2009) describe "dialectical bootstrapping" as a way to generate two estimates from an individual, with the hope of bracketing the true unknown value or quantity (Larrick and Soll 2006). When the unknown quantity is a probability of a binary event A, multiple estimates cannot bracket the true resolution of one or zero. One can elicit $P(A) \in [0,1]$ as the first judgment and $P(A^c) \in [0,1]$ for the second judgment, where A^c is the complement of the event A. These judgments can be logically considered two judgments for the same quantity $P(A)$, because they are linked using an axiom of probability $P(A) = 1 - P(A^c)$. However, in practice, the events A and A^c are not always linked logically in forecasters' cognition, and thus the judgments may be incoherent (Mandel 2008, 2005). Probabilistic incoherence can manifest in different ways because of refocusing or unpacking effects that are described in support theory (Tversky and Koehler 1994), but support theory cannot explain the incoherence of probability judgments of binary complements (Macchi et al. 1999).

Although the relationship between probabilistic coherence and accuracy has been described previously in the literature (Wright et al. 1994), its measurement and use within forecast aggregation has not been given full attention until recently (Tsai and Kirlik 2012, Wang et al. 2011). In addition to averaging $P(A)$ and $1-P(A^c)$ to produce a single judgment, we can also quantify the degree of incoherence as the difference between these values. Wang et al. (2011) were, to the best of our knowledge, the first to measure and weight individual forecasters according to their degree of probabilistic coherence. Rather than eliciting just two estimates and taking a simple average, they elicited dozens of estimates from each forecaster for related uncertain variables in the 2008 U.S. presidential election. The survey from which they took their data asked questions such as, "What is the probability that Obama wins Indiana?" and, "What is the probability that Obama wins Indiana and McCain wins Texas?"

Because a simple averaging of estimates is not well defined for many related, but logically nonequivalent events, Wang et al. (2011) applied the *coherent approximation principle* (CAP) (Predd et al. 2008, Osherson and Vardi 2006) to obtain a coherent set of probabilities that best represented the elicited incoherent subjective probabilities across all forecasters. The CAP was proposed to obtain a coherent set of forecast probabilities that is least different in terms of squared deviation from the elicited forecast probabilities. This "closest" set of coherent forecast probabilities is found by projecting the incoherent probabilities onto the coherent space of forecast probabilities, thus allowing incoherent probabilities to be approximated by coherent probabilities that are minimally different. The incoherence metric is then the Euclidean distance from an incoherent set of forecast probabilities to the "closest" coherent set of forecast probabilities.

For a simple event A, in terms of yielding an aggregated judgment for $P(A)$, the concept of the averaging $P(A)$ and $1-P(A^c)$ and the concept of the CAP are equivalent, and can be shown in Figure 1. For this example, an individual produces a judgment y_1 for $P(A)$ as 0.7, and then produces a judgment y_2 for $P(A^c)$ as 0.7. Taken together, these two judgments are incoherent and lie above the set of coherent estimates for $(P(A), P(A^c))$, which is represented by the line between $(0,1)$ and $(1,0)$. We can obtain an averaged estimate as $P(A)=(0.7+(1-0.7))/2=0.5$ and do the same to yield the averaged estimate of $P(A^c)=0.5$.

Figure 1 **The Coherent Approximation Principle (Predd et al. 2008, Osherson and Vardi 2006) for $P(A)$ and $P(A^c)$**



Equivalently, with the CAP, we can think of the coherent estimate $(0.5, 0.5)$ as the projection of $(0.7, 0.7)$ onto the set of coherent estimates, and we can uniquely measure the degree of incoherence as the Euclidean distance between the two points to yield an *incoherence metric* (IM). For this particular example, the IM is

$$\sqrt{(P(A)-y_1)^2+(P(A^c)-y_2)^2}$$
$$=\sqrt{(0.5-0.7)^2+(0.5-0.7)^2}.$$

Applying the CAP is a constrained optimization problem: minimization of the Euclidean distance, subject to constraints for the new probabilities to be coherent. Importantly, the CAP is an operational concept for *any* set of related subjective probabilities, given that the coherent set of probability forecasts can be expressed mathematically.

Wang et al. (2011) used a concept of *local coherence* to measure an individual judge's degree of coherence. After weighting each judge's distribution according to their local incoherence metric, giving less weight to more incoherent forecasters, the CAP was employed once within a complex numerical algorithm to find an aggregate coherent probability distribution that least modified the weighted average distribution. Conceptually, giving less weight to more incoherent forecasters should generate forecasting accuracy gains for multiple reasons. First, when a set of forecasts is incoherent, at least one estimate must be inaccurate (Mandel 2005). For every incoherent subjective probability distribution, there exists a coherent probability distribution that dominates the incoherent distribution in terms of the *Brier score*—a scoring rule described in more detail later (De Finetti 1990). Second, coherence might signal a more systematic accessing and consideration of relevant information. And third, coherence can indicate the care and effort taken by a forecaster to yield his or her estimate (Wang et al. 2011).

Initially, the optimization required for the CAP was equivalent to an NP-hard decision problem, but Predd et al. (2008) created an algorithm that decomposed the optimization into subproblems that used "local" sets of related events. The weighting scheme devised by Wang et al. (2011) maintained a similar run time, which was less than a full solution to the CAP but still substantial.

Using "big data" for the 2008 presidential election set, Wang et al. (2011) found that the coherent adjustments with more weight given to judges with greater individual coherence produced significantly greater stochastic accuracy—with upward of 41% improvement in average Brier score—and was comparable with other models like Intrade in predicting the outcomes of the United States in the election. However, this aggregation procedure was complex, and required significant computational expertise and time.

2.2. Novelty of This Paper

Coherence weighting within judgment aggregation is useful for one-time, unique events, where judges' previous performance is not known, or where a decision maker does not want to rely on perceived expertise and other demographics of judges, which have been shown to be poor indicators of forecasting performance (Burgman et al. 2011, Tetlock 2005). Although we incorporate the CAP and coherence weighting into the approach of this paper, we differ from Wang et al. (2011) in five key ways.

First, we are interested in how coherence weighting and "coherentizing" of probabilities can be used for one binary event A, rather than a large *given* collection of related probability events. Second, we ask what the best questions are and how many should we ask within our approach to increase the forecasting accuracy of the event A. Third, we consider simpler coherentizing algorithms that can be done without a need for significant computing time (see Wang et al. 2011) and expertise. We formulate a simple quadratic program that is done for each user-question pair, which implies the run time for the approach is linear in questions and users, rather than one that is severely increasing in these two factors. Fourth, when eliciting probabilities for multiple, unrelated, simple events, we consider that a judge might be an expert with respect to some events, and yet be uninformed for other events. Therefore, we weight individuals on each question rather than globally for all questions. And, fifth, we describe a different type of coherence-weighting function that has a natural threshold for decreasing the weight assigned to judges with extreme epistemic uncertainty.

Congruent aims of the paper are to assess the important elicitation-aggregation trade-off in efforts to boost forecasting accuracy. On the one hand, there is some evidence showing that aggregation of multiple judgments (e.g., judging the number of jellybeans in a jar) is improved when they are elicited independently from different forecasters (Surowiecki 2005). More recently, there is also evidence to suggest that multiple intraforecaster elicitations can improve accuracy, especially when the elicitations are spaced apart to promote independence. In one study, Vul and Pashler (2008) elicited two quantity judgments from individual forecasters for general knowledge questions, with the judgments elicited in immediate succession or separated by three weeks to promote independence. Averaging judgments in both cases improved accuracy, but the benefit was significantly greater in the spaced condition. As with interforecaster aggregation (crowdsourcing), intraforecaster aggregation seems to work best when the individual forecasts are not correlated, and spacing apart elicitations seems to promote that sort of independence.

On the other hand, some researchers proposed that judgment quality can be improved by eliciting judgments in ways that encourage a fuller consideration of pertinent information and the relation between alternative hypotheses (e.g., Hirt and Markman 1995, Sieck et al. 2007). In a classic study, Lord et al. (1984) demonstrated a correction to social judgment when they either made opposing possibilities more salient to people or directly instructed them to consider the opposite of their beliefs of the moment.

One way to promote a consideration of the relations between alternative hypotheses is to elicit estimates of each hypothesis in close succession or concurrently. For instance, Mandel (2005) found that probability judgment of A and A^c were more likely to be coherent (i.e., additive) if the judgments were elicited in immediate succession rather than spaced apart with an intervening distracter task. In a related vein, Williams and Mandel (2007) found an improvement in the quality of conditional probability judgments, both in terms of coherence (i.e., additivity) and accuracy (distance from mathematical probability), when queries were elicited with "evaluation frames" (a term coined by Tversky and Koehler 1994) that explicate the opposite possibility (e.g., "Given A, what is the probability of X rather than not-X?") rather than "economy frames" that only explicate a focal hypothesis (e.g., "Given A, what is the probability of X?"). As with the consecutive elicitations, evaluation frames encourage judges to think about one hypothesis in relation to other related hypotheses, increasing what Hsee (1996) calls their *evaluability*.

These two perspectives—improving aggregation by spacing apart elicitations and, improving individual judgments by eliciting probabilities together—present a trade-off that has yet to be carefully examined. Methods that encourage people to see the dependence between probabilities should improve judgment quality by increasing the accessibility of logical rules of probability and by improving the weighting of evidence among options. These advantages support an argument in favor of concurrent elicitation methods. However, methods that obscure the relations between probabilities through spaced elicitations should decrease coherence of related judgments and increase the variability of judgments and incoherence across forecasters. Thus, judgment aggregation procedures that capitalize on incoherence and variability should benefit from spaced elicitation methods. With our two studies described in the next section, we test both independent and concurrent elicitations in our aggregation approaches with the goal of obtaining the most accurate aggregate estimates.

3. Methods

3.1. Overview of the Two Studies

For a single event A, we want to determine the best way to elicit extra information from judges to use probabilistic incoherence to identify the "better" judges. Our ultimate goal is to maximize the improvement in judgment accuracy of $P(A)$, when compared with the equal-weighted averaging of judgments, by giving more weight to better judges. To simulate expert judgment with 60 simple A events, we constructed 60 statements, both true (e.g., $A = \{\text{Michelangelo painted the Sistine Chapel}\}$) and false (e.g., $A = \{\text{Melbourne is the capital of Australia}\}$), over general knowledge categories, and, in two studies, recruited undergraduate psychology student participants to serve as judges. Each participant provided confidence assessments rather than true forecasts. We asked the students about the veracity of each statement to generate $P(A)$. For example, if a participant believed a statement A was true with a confidence or subjective probability of 0.75, he or she was instructed to give $P(A) = 0.75$. To measure probabilistic incoherence, we also asked about the veracity of the related event statements B, with A and B mutually exclusive, A^c, and $A \cup B$.

The events A, B, A^c, and $A \cup B$ form a probability space of related events for each general knowledge category. In this research, A is the focal event, and the remaining three events serve as auxiliary events for the purpose of improving the forecasting accuracy of $P(A)$. In both studies, $P(B)$, $P(A^c)$, and $P(A \cup B)$ were elicited, but, recognizing that different elicitations may be more useful, the analyses within this paper were conducted using three alternative coherentization schemes: (a) the *two-way* (complements only) scheme relies on just $P(A)$ and $P(A^c)$; (b) the *three-way* (disjunctions only) scheme relies on $P(A)$, $P(B)$, and $P(A \cup B)$; and (c) the *four-way* (complements and disjunctions) scheme relies on all four probabilities.

Study 1 fostered independent intraperson judgments for the related events by spacing apart the related judgments in a manner that maximized inter-elicitation distance for related items (i.e., B, A^c, and $A \cup B$), whereas study 2 elicited the related judgments concurrently, thus minimizing inter-elicitation distance. This allowed us to test whether concurrently elicited probabilities improved the accuracy of the aggregated judgments and decreased the degree of incoherence. This also allowed comparison of the effects of coherence weighting of independent judgments versus coherence weighting of judgments that were assessed together.

In both studies, the statements used were the same, and a quadratic programming model was used to coherentize the estimates for each category and generate the incoherence metrics that were used to weight and aggregate the judgments for $P(A)$. The conjectures of the paper are described in testable hypotheses:

Hypothesis 1. *An equal-weighted average of coherentized estimates of $P(A)$ increases accuracy compared to an equal-weighted average of raw estimates of $P(A)$ (for both studies 1 and 2).*

Hypothesis 2. *A coherence-weighted average of coherentized estimates of $P(A)$ increases accuracy compared to an equal-weighted average of coherentized estimates of $P(A)$ (for both studies 1 and 2).*

Hypothesis 3. *An equal-weighted average of raw estimates of $P(A)$ is more accurate when related estimates (i.e., $B, A^c, A \cup B$) are elicited concurrently rather than in a spaced manner.*

Hypothesis 4. *A coherence-weighted average of coherentized $P(A)$ is more accurate when related estimates are elicited in a spaced manner rather than concurrently.*

Hypotheses 1 and 2 are formulated with Wang et al. (2011) in mind, and Hypothesis 3 is formulated with Sieck et al. (2007), Mandel (2005), Hsee (1996), Hirt and Markman (1995), and Lord et al. (1984) in mind. We assume that spaced estimates will be more incoherent than concurrent estimates. For Hypothesis 4, we further assume that our method will take full advantage of the added incoherence. We examine the initial hypotheses (Hypotheses 1 and 2) for the two-way, three-way, and four-way coherentization schemes, and we examine Hypothesis 4 for the optimal scheme and close competitors.

3.2. Experiment Design

Both studies 1 and 2 used the same 60 general knowledge categories, where each category contained statements $A, B, A^c, A \cup B$. Asking for a probability estimate concerning the veracity of each statement generated a 240-question survey. The statements used in the studies were designed such that freshmen undergraduate psychology students would have familiarity with at least some of the topics. The statements were therefore spread over topics in history, geography, psychology, economics, postal abbreviations, state/country capitals, art, politics, science, sports, and other topics.

An example A statement was "In the Earth's solar system, Mars is the fifth plant from the Sun," which is false. In this category, the B statement was "In the Earth's solar system, Jupiter is the fifth planet from the Sun," the A^c statement was "In the Earth's solar system, Mars is NOT the fifth planet from the Sun," and $A \cup B$ statement was "In the Earth's solar system, Mars or Jupiter is the fifth planet from the sun." Additional examples of statements are found in Appendix A.

In both studies, student participants, who were blind to the full purpose of the research, served as judges and provided the probabilities to be aggregated. The student participants were only incentivized to participate with credit/no credit for fulfilling a course requirement, and this decision was based only on completion, and not performance. Each student participant was presented with a statement, and used a pull-down menu to select his or her subjective probability from 0% to 100% that the statement was true. The elicited probability was then displayed on a ruler beneath the estimate to give the participant a visual aid. The participant then submitted an answer and moved to the next screen with the next statement(s).

3.3. Coherentization

Letting the vector y be the elicited vector of subjective probability estimates for a coherentization scheme, we have p as the "closest" coherent vector of probabilities. Finding p is done by coherentization using a quadratic programming model.[1] For example, for the four-way scheme, $y = [y_1, y_2, y_3, y_4]$ is elicited, and $p = [P(A), P(B), P(A^c), P(A \cup B)]$ is found as

$$\text{Minimize (over } p\text{): } (y_1 - P(A))^2 + (y_2 - P(B))^2 + \big(y_3 - P(A^c)\big)^2 + \big(y_4 - P(A \cup B)\big)^2,$$

[1] We use the standard quadratic programming package in MATLAB, called "quadprog."

such that

1. $P(A)+P(A^c)=1$,
2. $P(A)+P(B)=P(A\cup B)$ (since A and B are mutually exclusive),
3. $0\le P(A), P(B), P(A\cup B),\ P(A^c)\le 1$.

In all elicitation cases, the objective function is the squared Euclidean distance between y and p. The square root of the objective function yields the incoherence metric for the category for each participant. In the example of the four-way scheme, the first two constraints are represented as linear equality constraints, and the last set of constraints are represented as linear inequality constraints.[2] Taken together, the three sets of coherence constraints form a *convex* set of coherent probabilities. The convexity of the set of coherent probabilities implies that any weighted average of coherent probabilities is again coherent, as a weighted average is a convex combination of points.

As an example of the two-, three-, and four-way coherentization schemes, consider the incoherent estimates for $y=[y_1,y_2,y_3,y_4]=[0.4,0.3,0.5,0.6]$, where y_1 is an estimate of $P(A)$, y_2 is an estimate for $P(B)$, y_3 is an estimate for $P(A^c)$, and y_4 is an estimate for $P(A\cup B)$. Then the multiple formulations yield coherentized probabilities, such as:

- Two-way: $[P_c(A),P_c(A^c)]=[0.45,0.55], \mathrm{IM}=0.07$.
- Three-way: $[P_c(A),P_c(B),P_c(A\cup B)]=[0.37,\ 0.27,\ 0.63]$, $\mathrm{IM}=0.06$.
- Four-way: $[P_c(A),P_c(B),P_c(A^c),P_c(A\cup B)]=[0.42,\ 0.24,\ 0.58,\ 0.66]$, $\mathrm{IM}=0.12$.

3.4. Aggregating Probabilities

For aggregating, given y_1^i as the raw subjective probability for $P(A)$ from the ith student participant, and $P_c^i(A)$ as the coherentized probability for $P(A)$ from the ith participant, the simple equal-weighted average of the raw estimates for all N participants is defined as

$$\frac{1}{N}\sum_{i=1}^{N} y_1^i;$$

the simple, equal-weighted average of the coherentized estimates is similarly defined as

$$\frac{1}{N}\sum_{i=1}^{N} P_c^i(A);$$

and the coherence-weighted average of coherentized estimates is defined as

$$\frac{1}{\lambda}\sum_{i=1}^{N}\omega(\mathrm{IM}^i)P_c^i(A), \quad \text{with } \lambda=\sum_{i=1}^{N}\omega(\mathrm{IM}^i),$$

where $\omega(\mathrm{IM}^i)$ is the weighting function evaluated at the ith participant's incoherence metric for the category containing A.

When weighting forecasters according to IM^i, with IM^i defined as the Euclidean distance between y^i and p^i, the weighting function is defined on the nonnegative real line, and the function decreases as IM^i increases to increasingly penalize incoherent forecasts. Because the weights are normalized during the aggregation, only the ratio values of the weights are relevant (not the absolute values). Arbitrarily setting the weighting for a perfectly coherent set of forecasts ($\mathrm{IM}^i=0$) to a value of 1, a set of weighting functions that satisfy these conditions is described as

$$\omega(\mathrm{IM}^i)=\left(\frac{\mathrm{IM}_{\max}-\mathrm{IM}^i}{\mathrm{IM}_{\max}}\right)^{\beta},$$

where $\mathrm{IM}_{\max}$ is the largest incoherence score recorded for any category over all participants and all questions. The value of $\mathrm{IM}_{\max}$ is always known before resolution, and all IM^i values are assigned nonnegative weights. The parameter β is a scale parameter, and when $\beta=0$, all IM values receive a weight value of one, thus reducing to a simple equal-weighted averaging. When $\beta=1$, the weighting function is a linear, decreasing function that weights perfect coherence as a value of one, and the largest incoherence metric as zero. As β approaches infinity, only the perfectly coherent forecasts ($\mathrm{IM}^i=0$) are assigned a nonzero weight.

Given that the coherentizing and coherence-weighting approaches are performed before the resolutions of the questions are known, an important question concerns a best value for the β parameter. Previous literature provides a good first guess. In particular, a significant challenge when eliciting probabilities for a diverse set of events is that judges commonly use the probability value of 0.5 to indicate a probability value of 0.5 (i.e., a point-mass value at 0.5 for a fair coin flip) but also use 0.5 to indicate epistemic uncertainty (i.e., a uniform 0–1 distribution). In the first case, the forecaster is describing aleatory uncertainty (Pate-Cornell 1996), and has assessed that

[2] If one were to ask conditional probability questions, the constraints would no longer be completely linear and would include ratio constraints.

both the events A and A^c are equally likely. However, in the second case, the forecaster might not have sufficient knowledge of the event space or other information to make an informed forecast.

Although conceptually one could argue that 0.5 is the appropriate point estimate for epistemic uncertainty, the excess of 0.5 probabilities can lessen the influence of judges that might have important insight when an equal-weighted average is used to produce an aggregate estimate. The epistemic use of 0.5 implies a blip or jump at 0.5 in the histogram of elicited probabilities (Bruine de Bruin et al. 2002). A particular challenge for forecast aggregation is then to differentiate the forecasters that are describing aleatory uncertainty, and those that are describing epistemic uncertainty.

When there is a significant degree of epistemic uncertainty among judges, some will likely enter 0.5 for $P(A)$, $P(B)$, and $P(A \cup B)$ (especially if the estimates are independently elicited), which will yield incoherence because $P(A)+P(B) \neq P(A \cup B)$. Thus the IM score can be advantageous for the three-way and four-way coherentization schemes because a response of 0.5 to express epistemic uncertainty will yield an IM value of approximately 0.29 for the three-way scheme and 0.32 for the four-way scheme, but a response of 0.5 to express aleatory uncertainty (i.e., $P(A) = P(B) = 0.5$ and $P(A \cup B) = 1$) will yield an IM value of zero. We therefore want to pick a β value appropriately to assign sufficiently small weights to these epistemic judgments and judgments that are further away from the coherent set of judgments.

With this in mind, the scale parameter is fixed for the following studies, $\beta = 15$, which yields significantly small weight values of $\omega(0.29) = 0.021$, and $\omega(0.32) = 0.013$. In a later section, we demonstrate how this parameter is sufficient to alleviate issues with the fifty–fifty blip, and also demonstrate the robustness of our gains in forecasting accuracy using sensitivity analysis.

4. Experiments

4.1. Study 1: Spaced Judgments of Related Probabilities

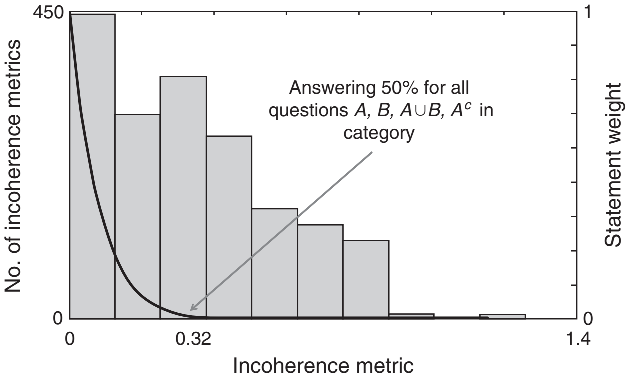
Study 1 featured 30 undergraduate George Mason University psychology students who provided the probability estimates to be aggregated. As noted earlier, the important distinction between the first and second studies was that the probabilities for events in the same category were spaced as far apart as possible in study 1 to try to foster *independent intraparticipant judgments* for each category, whereas the probabilities for events in the same category were elicited concurrently in study 2. In study 2, all statements in a category received probability judgments together, although the statement order was randomized. In study 1, the first randomly chosen statement for a category received a probability judgment, but the next statement in the same category did not receive a judgment until the participant cycled through unrelated statements of the other 59 categories. With the randomization, each participant did not know if they were providing a probability for A, B, A^c, or $A \cup B$, and the participant was not allowed to change previously submitted probabilities.

Instructions were to enter “a probability between 0% and 100%. If you are absolutely certain that the statement is true, you should enter 100. Likewise, if you are absolutely certain that the statement is false, you should enter 0. If you are uncertain, you should enter the probability that corresponds with what you think are the chances that the statement is true.” Participants took an average of 45 minutes to submit all answers.

After all surveys were completed, the IM scores were calculated for each judge-category pair, and Hypotheses 1 and 2 were tested for the two-way, three-way, and four-way coherentization schemes. Figure 2 shows the weighting function (right axis) transposed over the histogram (left axis) of incoherence metrics for the 1,800 participant-category pairs for the four-way coherentization. We see from this figure that a majority of participants were beyond the 0.32 cutoff that resulted from answering 0.5 for $P(A), P(B), P(A^c)$, and $P(A \cup B)$. We observed similar results with the three-way coherentization.

Figure 3 shows the histogram of all elicited probabilities y_1^i for $P(A)$ in the top panel, and shows in the bottom panel the 861 coherentized $P_c^i(A)$ estimates that received IM^i scores less than or equal to 0.31 for the four-way coherentization. We note the histogram bar that included 0.5 decreased the most from the top to the bottom panel. This observation supports the

Figure 2 **The Weighting Function with Scale Parameter Set as $\beta = 15$ (Right Axis), and Histogram of the Incoherence Metrics for Study 1 (Left Axis)**



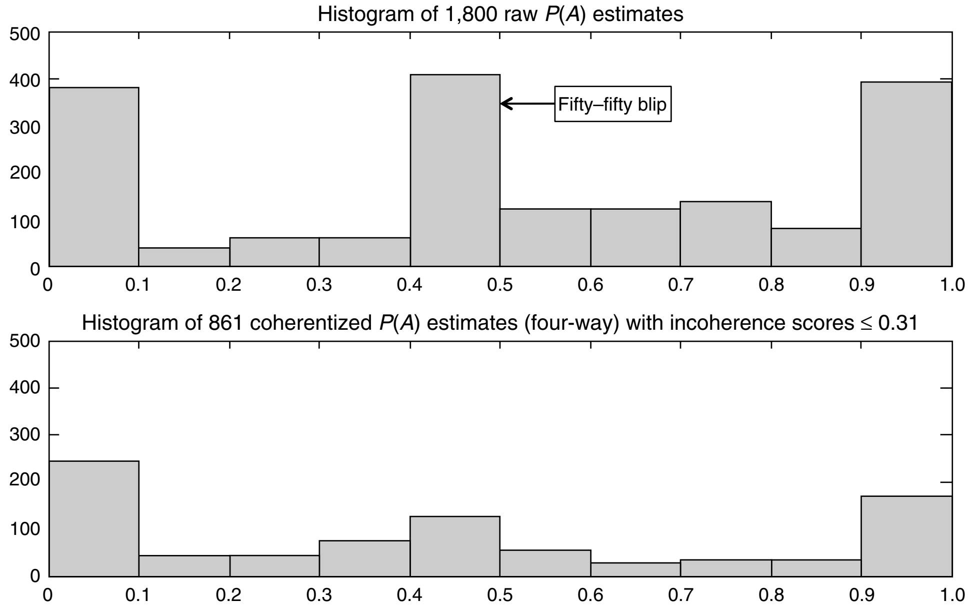
Note. The weighting function assigns to each incoherence metric a weight that is normalized and used for aggregating over the participants.

conjecture that the weighting approach would reduce the fifty–fifty blip.

We used two tests for Hypotheses 1 and 2. First, after we generated all aggregate estimates for $P(A)$ for all 60 categories, we looked at an average Brier score (Brier 1950). The Brier score is a proper scoring rule for judgment accuracy, which may be further decomposed to provide measures of calibration and discrimination (Yaniv et al. 1991, Murphy 1973), and the Brier score is one of the most popular scoring rules (Gneiting 2011). However, for statistical tests we looked at the number of times (out of 60) that each aggregation approach produced an estimate that was closer to the correct resolution value, and we looked at the change in average absolute distance between estimates and resolution values. All statistical tests in this paper are unidirectional, so p-values are reported for one-tailed tests as consistent with our hypotheses. For some hypotheses, multiple tests must be run, requiring Šidák correction ($p < 0.017$ for three t-tests).

Table 1 displays the average Brier score over the 60 questions for the various aggregation approaches. In these cases, the Brier score for a binary event scores an elicited probability forecast as BS(forecast) = (resolution-forecast)^2, where *resolution* is the value 0 if the statement is false, and 1 if it is true. The first row of the table displays the average Brier score of the simple, equal-weighted average of the raw estimates y_1^i. The next three rows display the equal-weighted average of the coherentized judgments $P_c^i(A)$ for the various coherentization schemes. We see that the two-way scheme does not yield much improvement over the LINOP. The three-way scheme

Figure 3 **The Histogram of 1,800 Raw Estimates y_1^i (Top), and Histogram of 861 Coherentized Estimates $P_c^i(A)$ for Categories with Incoherence Metrics Less Than 0.31 (Bottom) for Study 1**



Note. Coherentization most affected the bin containing probability estimates of 50%.

Table 1 The Average Brier Scores for the Various Aggregation Approaches for $P(A)$, and the Percent-Improvement Over the Equal-Weighted Averaging of Raw Estimates y_1^i (*LINOP*) for Study 1

Aggregation approach	Average BS (0–1 scale)	Improvement over *LINOP* (%)
Equal-weighted estimate just using raw y_1^i (*LINOP*)	0.2243	—
Equal-weighted estimate using two-way $P_c^i(A)$, $P_c^i(A^c)$	0.2218	1.14
Equal-weighted estimate using three-way $P_c^i(A)$, $P_c^i(B)$, $P_c^i(A\cup B)$	0.1831	18.39
Equal-weighted estimate using four-way $P_c^i(A)$, $P_c^i(B)$, $P_c^i(A^c)$, $P_c^i(A\cup B)$	0.1932	13.86
Coherence-weighted estimate using two-way $P_c^i(A)$, $P_c^i(A^c)$	0.2066	7.88
Coherence-weighted estimate using three-way $P_c^i(A)$, $P_c^i(B)$, $P_c^i(A\cup B)$	0.1515	32.45
Coherence-weighted estimate using four-way $P_c^i(A)$, $P_c^i(B)$, $P_c^i(A^c)$, $P_c^i(A\cup B)$	0.1568	30.10

Table 2 The Number of Times the Equal-Weighted Averaging of Coherentized Estimates $P_c^i(A)$ Improved Over the Equal-Weighted Averaging of Raw Estimates y_1^i (*LINOP*), and the Average Absolute Improvement Distance and 90% Confidence Intervals for Study 1

Aggregation approach	No. of times improved over *LINOP* (out of 60)	Average absolute improvement, 90% CI
Equal-weighted estimate using two-way $P_c^i(A)$, $P_c^i(A^c)$	28	0.0006, [−0.0081, 0.0093]
Equal-weighted estimate using three-way $P_c^i(A)$, $P_c^i(\mathrm{B})$, $P_c^i(A\cup B)$	41*	0.0448, [0.0177, 0.0718]
Equal-weighted estimate using four-way $P_c^i(A)$, $P_c^i(\mathrm{B})$, $P_c^i(A^c)$, $P_c^i(A\cup B)$	41*	0.0296, [0.0087, 0.0506]

*Statistically greater than 30 with $\alpha = 0.05$.

offers about 18% improvement, which is greater than the four-way scheme. In the bottom three rows, we see that the coherence weighting of the coherentized estimates offers an additional improvement over the equal-weighted average of the coherentized estimates, and the three-way and the four-way coherentization schemes offer about 30% total improvement over the *LINOP*. Using the average Brier score, we see evidence that supports both Hypotheses 1 and 2 for the three- and four-way coherentizations.

A formal test of the first two hypotheses looks at the number of questions in which the respective methods offered an improvement over the baseline method. For testing Hypothesis 1, we refer to Table 2, and see that the equal-weighted averages of coherentized estimates for the three-way and the four-way schemes produced a large proportion of estimates closer to the correct resolution when compared to the equal-weighted averages of the raw estimates, but the two-way scheme did not. Using the normal approximation with the sample proportion, we show that the proportion of questions (41 out of 60) for which the three- and four-way coherentization schemes showed improvement was significantly greater than 0.5 (three-way and four-way: $t_{59} = 3.053$, $p = 0.002$), which is the proportion one would expect if neither method were superior. In Table 2, we also see that the three- and four-way coherentization schemes move the probabilities on average about 0.045 and 0.030 closer to the correct resolution, respectively, gains that are statistically greater than zero (three-way: $t_{59} = 2.766$, $p = 0.004$; four-way $t_{59} = 2.361$, $p = 0.011$). Table 2 also displays the respective 90% confidence intervals.

For testing Hypothesis 2, we refer to Table 3, and see that a coherence-weighted average of coherentized probabilities gives better estimates when compared to the respective methods that use an equal-weighted average of the coherentized estimates. In this case, all three weighting methods produced proportions of improvement over the equal-weighted average of the coherentized estimates that are greater than 0.5 (two-way: $t_{59} = 4.884$, $p < 0.001$; three-way: $t_{59} = 4.472$, $p < 0.001$; four-way: $t_{59} = 6.339$, $p < 0.001$). The average distance of improvement is also greater with coherence weighting than with coherentizing alone, and these distances are statistically greater than zero for all three cases (two-way: $t_{59} = 4.084$, $p < 0.001$; three-way: $t_{59} = 5.699$, $p < 0.001$; four-way: $t_{59} = 5.365$, $p < 0.001$). We see that, in this case, the four-way coherentization offers the greatest average improvement distance, but the three-way scheme is close behind. Šidák corrections do not change any of these conclusions.

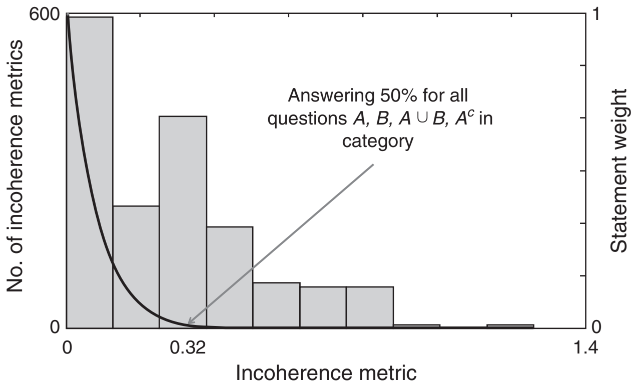
The results of this study indicate that the three-way elicitation of event probabilities may be the best elicitation approach for subsequently implementing coherentization and coherence weighting. It requires one less elicitation per question than the four-way scheme, while yielding similar improvement. Both the three- and four-way coherentization schemes dominate the two-way scheme for improving accuracy.

Table 3 The Number of Times the Coherence-Weighted Averaging of Coherentized Estimates Improved $P_c^i(A)$ Over the Equal-Weighted Averaging of Coherentized Estimates $P_c^i(A)$, and the Average Absolute Improvement Distance and 90% Confidence Intervals for Study 1

Aggregation approach	No. of times improved over equal-weighted, coherent (out of 60)	Average absolute improvement, 90% CI
Coherence-weighted estimate using two-way $P_c^i(A)$, $P_c^i(A^c)$	46*	0.0401, [0.0237, 0.0566]
Coherence-weighted estimate using three-way $P_c^i(A)$, $P_c^i(B)$, $P_c^i(A \cup B)$	45*	0.0736, [0.0520, 0.0951]
Coherence-weighted estimate using four-way $P_c^i(A)$, $P_c^i(B)$, $P_c^i(A^c)$, $P_c^i(A \cup B)$	49*	0.0959, [0.0660, 0.1257]

*Statistically greater than 30 with $\alpha = 0.05$.

Figure 4 The Weighting Function with Scale Parameter Set as $\beta = 15$ (Right Axis), and Histogram of the Incoherence Metrics for Study 2 (Left Axis)



Note. The weighting function assigns to each incoherence metric a weight that is normalized and used for aggregating over the participants.

4.2. Study 2: Concurrent Judgments of Related Probabilities

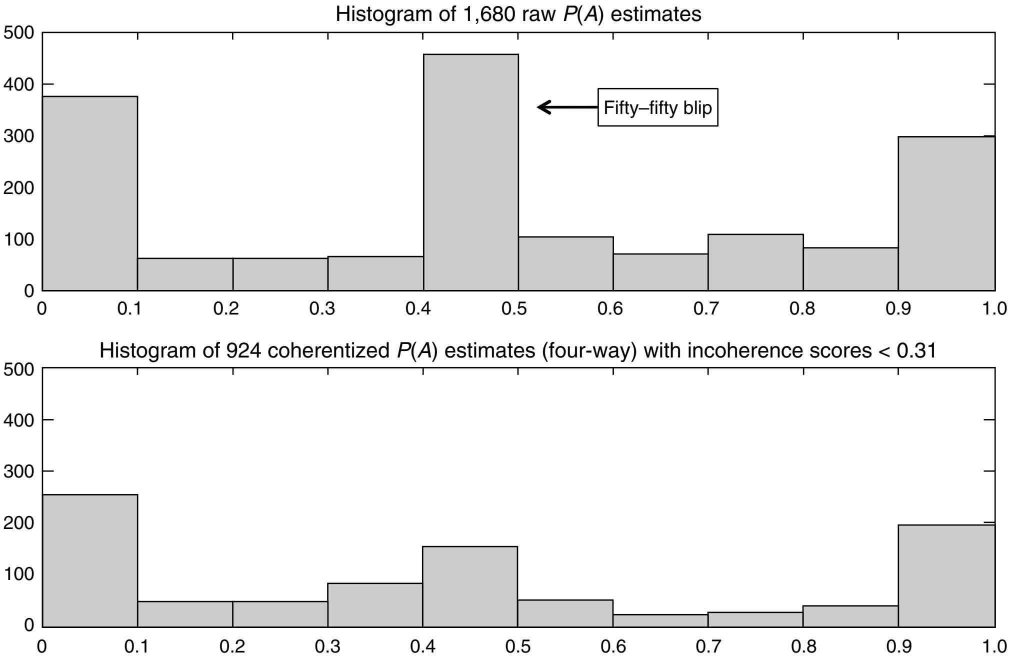
Study 2 featured a different sample of 28 undergraduate GMU psychology students who provided the probability estimates to be aggregated. The aims of study 2 were to measure the effect of concurrently elicited related probability judgments for coherentization and coherence weighting for comparison with study 1. Overall, study 2 had the same setup as study 1, except estimates for $P(A)$, $P(B)$, $P(A \cup B)$, $P(A^c)$ were elicited together on the same screen rather than spaced apart.

We begin with an examination of the effect of coherentization and coherence weighting on judgment accuracy and turn to a test of our cross-study comparisons pertinent to testing Hypotheses 3 and 4 in §4.3. Figure 4 is similar to Figure 2 and displays the weighting function (right axis) overlaid on the histogram of incoherence metrics. We note that there are more forecasts in the left-most coherent bar for study 2, but this is still a large degree of incoherence over all the estimates. Figure 5 is similar to Figure 3, and shows the histogram of all elicited probabilities y_1^i in the top panel, and the bottom panel shows the 924 coherentized $P_c^i(A)$ estimates that received IM scores less than or equal to 0.31 for the four-way elicitation. We note that although we elicited fewer total estimates from participants in study 2, we had more estimates with incoherence metrics less than or equal to 0.31. We also note similar support for the conjecture that the weighting approach reduced the fifty–fifty blip.

Table 4 displays the average Brier score over the 60 questions for the various aggregation approaches for study 2, and displays in parentheses the average Brier score for study 1. The first row displays the average Brier score of the simple, equal-weighted average of the raw estimates y_1^i. The next three rows display the equal-weighted average of the coherentized forecasts, $P_c^i(A)$. As with study 1, we see that the two-way coherentization scheme does not yield much improvement over the equal-weighted average of the raw estimates. The three-way scheme offers greater improvement than the four-way scheme. We see that the coherence weighting of the coherentized estimates offers an improvement over the equal-weighted average of the coherentized estimates for study 2, and for these cases, the three-way and the four-way schemes are very close.

As with study 1, we formally test Hypotheses 1 and 2 for study 2 by looking at the number of questions that the respective methods offered an improvement over the compared method. For testing Hypothesis 1, we refer to Table 5, which shows that the equal-weighted averages of coherentized estimates for the three- and four-way schemes offer a proportion of improvement over the equal-weighted average of the raw estimates that is statistically significant (three-way: $t_{59} = 3.053$, $p = 0.002$; four-way: $t_{59} = 3.381$, $p = 0.001$). We see that the three- and

Figure 5 The Histogram of 1,680 Raw Estimates y_1^i (Top), and Histogram 924 of Coherentized Estimates $P_c^i(A)$ for Categories with Incoherence Metrics Less Than 0.31 (Bottom) for Study 2



Note. We note the reduction in estimates at the probability value 0.5.

four-way coherentization schemes move the probabilities, on average, about 0.031, and 0.015 closer to the correct resolution, respectively. Because we are making three comparisons to test Hypothesis 1, we must control for family-wise alpha. Using our stricter standard ($p < 0.017$), we see that gains for the three-way scheme are statistically significant, but gains for the four-way scheme are not (three-way: $t_{59} = 2.462$, $p = 0.008$; four-way: $t_{59} = 2.009$, $p = 0.025$).

For testing Hypothesis 2, we refer to Table 6, and see that a coherence-weighted average of coherentized

Table 4 The Average Brier Scores for the Various Aggregation Approaches for $P(A)$, and the Percent-Improvement Over the Equal-Weighted Averaging of Raw Estimates y_1^i (*LINOP*) for Study 2

Aggregation approach	Average BS (0–1 scale), (study 1 BS)	Percent-improvement over *LINOP* (%)
Equal-weighted estimate just using raw y_1^i (*LINOP*)	0.2020 (0.2243)	—
Equal-weighted estimate using two-way $P_c^i(A), P_c^i(A^c)$	0.2076 (0.2218)	−2.77
Equal-weighted estimate using three-way $P_c^i(A), P_c^i(B), P_c^i(A \cup B)$	0.1778 (0.1831)	11.96
Equal-weighted estimate using four-way $P_c^i(A), P_c^i(B), P_c^i(A^c), P_c^i(A \cup B)$	0.1865 (0.1932)	7.68
Coherence-weighted estimate using two-way $P_c^i(A), P_c^i(A^c)$	0.1928 (0.2066)	4.54
Coherence-weighted estimate using three-way $P_c^i(A), P_c^i(B), P_c^i(A \cup B)$	0.1704 (0.1515)	15.65
Coherence-weighted estimate using four-way $P_c^i(A), P_c^i(B), P_c^i(A^c), P_c^i(A \cup B)$	0.1708 (0.1568)	15.45

Table 5 The Number of Times the Equal-Weighted Averaging of Coherentized Estimates $P_c^i(A)$ Improved Over the Equal-Weighted Averaging of Raw Estimates of y_1^i (*LINOP*), and the Average Absolute Improvement Value and 90% Confidence Intervals for Study 2

Aggregation approach	No. of times improved over *LINOP* (out of 60)	Average absolute improvement, 90% CI
Equal-weighted estimate using two-way $P_c^i(A), P_c^i(A^c)$	20	−0.0072, [−0.0138, 0.0007]
Equal-weighted estimate using three-way $P_c^i(A), P_c^i(B), P_c^i(A \cup B)$	41*	0.0311, [0.0100, 0.0522]
Equal-weighted estimate using four-way $P_c^i(A), P_c^i(B), P_c^i(A^c), P_c^i(A \cup B)$	42*	0.0149, [0.0025, 0.0273]

*Statistically greater than 30 with $\alpha = 0.05$.

Table 6 The Number of Times the Coherence-Weighted Averaging of Coherentized Estimates $P_c^i(A)$ Improved Over the Equal-Weighted Averaging of Coherentized Estimates $P_c^i(A)$, and the Average Absolute Improvement Value and 90% Confidence Intervals for Study 2

Aggregation approach	No. of times improved over equal-weighted, coherent (out of 60)	Average absolute improvement, 90% CI
Coherence-weighted estimate using two-way $P_c^i(A), P_c^i(A^c)$	43*	0.0274, [0.0181, 0.0367]
Coherence-weighted estimate using three-way $P_c^i(A), P_c^i(B), P_c^i(A \cup B)$	38*	0.0286, [0.0074, 0.0498]
Coherence-weighted estimate using four-way $P_c^i(A), P_c^i(B), P_c^i(A^c), P_c^i(A \cup B)$	43*	0.0521, [0.0267, 0.0775]

*Statistically greater than 30 with $\alpha = 0.05$.

probabilities gives better estimates when compared to the respective methods that use an equal-weighted averaging of the coherentized estimates. In this case, all three weighting methods produce proportions of improvement that are greater than 0.5 (two-way: $t_{59} = 3.724$, $p < 0.001$; three-way: $t_{59} = 2.143$, $p = 0.018$; four-way: $t_{59} = 3.724$, $p < 0.001$). Tests were significant for all but the three-way scheme, given the conservative family-wise alpha standard for Šidák correction to our tests. The average distance of improvement is also greater with coherence weighting than with coherentizing alone, and these distances are statistically greater than zero for all three cases (two-way: $t_{59} = 4.915$, $p < 0.001$; three-way: $t_{59} = 2.258$, $p = 0.014$; four-way: $t_{59} = 3.431$, $p < 0.001$). Again we see that three-way scheme offers comparable improvement to the four-way scheme, and given that the three-way requires one less elicitation, we view this as our best method going forward.

4.3. Comparing Results of Studies 1 and 2

Given studies 1 and 2, we can explicitly test the effect of having independent intraparticipant judgments for the subjective probabilities in each category. Figures 2 and 4 allow us to compare the histograms of incoherence metrics. In general, and as predicted, there were more coherent estimates in study 2 than in study 1. The leftmost histogram bar in study 2 (Figure 4) contains almost 600 estimates, whereas the same bar in study 1 (Figure 2) contains about 450 estimates. In study 1, the 95% confidence interval for the average IM value for the four-way scheme was [0.3208, 0.3441] and [0.2294, 0.2491] for the three-way scheme, whereas the 95% confidence interval for the average IM in study 2 for the four-way scheme was [0.2559, 0.2784] and [0.1773, 0.1952] for the three-way scheme.

We test Hypothesis 3 to see the improvement in the equal-weighted average of raw estimates of $P(A)$ (*LINOP*) when the estimates of $P(A)$, $P(B)$, $P(A^c)$, and $P(A \cup B)$ are provided concurrently rather than in the spaced manner. In Table 4, we see that the average Brier score of the raw estimates y_1^i (*LINOP*) in study 2 is less than that of study 1. In Table 7, for each method, we compared the number of questions where the aggregate estimate of study 1 was better than the aggregate estimate of study 2, and we examined the distance between the averages. For the equal-weighted estimates of using the raw y_1^i, the proportion of times that the estimate of study 1 beat the estimate of study 2 was 0.367 (22/60), which is statistically less than 0.5 ($t_{59} = 2.143$, $p = 0.018$), and the average distance between the question estimates is less than zero ($t_{59} = 1.763$, $p = 0.042$). These findings support our prediction that participants would make more accurate raw estimates when their related estimates were elicited concurrently in study 2 rather than spaced independently as in study 1. We also see some support for the equal-weighted estimate of coherentized probabilities of study 2 being slightly better when we look at the average Brier scores in Table 4.

We test Hypothesis 4 to see the improvement in the coherence-weighted average of coherentized estimates of $P(A)$ when the estimates of $P(A)$, $P(B)$, $P(A^c)$, and $P(A \cup B)$ are provided in the spaced manner, rather than concurrently. We test the three-way scheme, deemed the best performing method, for Hypothesis 4, and in Table 4, we see that the average Brier score for the three-way coherence weighting scheme in study 1 is less than that in study 2. In testing Hypothesis 4 for the three-way scheme, for the coherence weighting of coherentized estimates, we see that the proportion of questions for which study 1 produces a better estimate than study 2 is greater than 0.5 ($t_{59} = 1.858$, $p = 0.034$). Furthermore, the average absolute improvement distance is greater than

Table 7 **The Number of Times the Estimate of Study 1 Was Closer to the Resolution Than the Estimate of Study 2 for the Various Aggregation Approaches**

Aggregation approach	No. of times study 1 better than study 2	Average absolute improvement of study 1 over study 2, 90% CI
Equal-weighted estimate just using raw y_1^i (*LINOP*)	22*	−0.021, [−0.0408, −0.0011]
Equal-weighted estimate using two-way $P_c^i(A)$, $P_c^i(A^c)$	28	−0.0131, [−0.0334, 0.0072]
Equal-weighted estimate using three-way $P_c^i(A)$, $P_c^i(B)$, $P_c^i(A\cup B)$	28	−0.0073, [−0.0217, 0.0071]
Equal-weighted estimate using four-way $P_c^i(A)$, $P_c^i(B)$, $P_c^i(A^c)$, $P_c^i(A\cup B)$	29	−0.0062, [−0.023, 0.0105]
Coherence-weighted estimate using two-way $P_c^i(A)$, $P_c^i(A^c)$	27	−0.0003, [−0.0335, 0.0329]
Coherence-weighted estimate using three-way $P_c^i(A)$, $P_c^i(B)$, $P_c^i(A\cup B)$	37**	0.0376, [0.0021, 0.0732]
Coherence-weighted estimate using four-way $P_c^i(A)$, $P_c^i(B)$, $P_c^i(A^c)$, $P_c^i(A\cup B)$	37**	0.0375, [−0.0096, 0.0846]

Note. The average improvement and 90% confidence intervals are also shown.

*Statistically less than 30 with $\alpha = 0.05$.

**Statistically greater than 30 with $\alpha = 0.05$.

zero for the three-way elicitation (three-way: $t_{59} = 1.771$, $p = 0.041$).

In sum, although we do see a slight expected increase in accuracy of the equal-weighted average of the raw estimates by showing the participants the related judgments on the same screen, we have generated larger gains in accuracy by using coherence weighting. By spacing out the related judgments in study 1, we increased the degree of incoherence among the estimates when compared to study 2, but within the three-way coherence weighting scheme, we see independent intraparticipant judgments provide the best estimates of any of the tested approaches.

5. Discussion

5.1. Summary of Findings

The key insights from the two studies are (i) concurrent judgments improved the raw estimate of $P(A)$ and reduced the degree of incoherence, although they did not eliminate it, (ii) coherentizing individuals' judgments improved accuracy, (iii) coherence weighting generated larger accuracy gains than coherentizing alone, and (iv) independence across related intraparticipant judgments was used to generate more accurate aggregate forecasts with coherence weighting.

Table 8 **Summarizing the Effects in Study 1 of Three Aggregation Approaches, with Both Raw and Coherentized Estimates of** $P(A)$**, on the Average Brier Score Using the Three-Way Coherence Scheme**

	Aggregation approach			
Probabilities	Equal weighting	Coherence weighting	Equal weighting of top *n*-most coherent forecasters for each question	*n*
Raw y_1^i	0.2243	0.1539	0.1699	3
			0.1625	5
			0.1560	10
			0.1731	15
Coherentized $P_c^i(A)$	0.1831	0.1515	0.1685	3
			0.1617	5
			0.1533	10
			0.1653	15

Note. The threshold approach in the two rightmost columns of the table shows the average Brier scores when only the top *n*-most judges (out of 30) are aggregated with equal weight.

Overall, when looking at all of the aggregation approaches tested in terms of average Brier score, the most accurate method was the three-way coherence weighting of the coherentized estimates that were elicited with maximum independence in the spaced mode (study 1). In our two studies, separate elicitations of related probabilities generated less accurate initial estimates of the target probability than did joint elicitations, allowing more room for improvement by coherentization. Likewise, less coherent estimates allowed greater improvement by coherence weighting.

The three-way scheme performed best, and from it, the effects of both coherentization and coherence weighting are shown in terms of the average Brier score in Table 8. We see that coherentizing probabilities alone provides some increase in accuracy, but coherence weighting provides the largest gains in accuracy for both raw and coherentized probabilities. In coherence weighting, the least coherent probability estimates are weighted significantly less than the most coherent estimates. Thus, the effect of coherentization is minimal with the coherence weighting.

In Table 8, we also examine how threshold weighting (Tsai and Kirlik 2012) compares with our weighting function approach. In the rightmost columns of Table 8, we show the Brier scores for equal weighting of the *n-most* coherent judges, for various values of n (out of 30). We see the most accurate aggregate estimates are obtained with an n of about 10, and these methods produce similar gains in accuracy with the three-way method described within this paper. However, we note that the inverse of the coherence metric employed by Tsai and Kirlik is not the same as our IM, as their measures describe the coherence of a judge's estimates with Bayes' theorem and historical data.

5.2. Sensitivity Analysis

The major efforts for sensitivity analysis concern the weighting function employed and the pooling size. In particular, we found that our Brier score results and comparisons between the two studies were generally robust to different weighting functions, provided that the weighting functions assigned very little weight (< 0.05) to the critical threshold point that represented answering 0.5 for all questions in the category.

Figure 6 shows in the top panel the average of the BS for the three-way scheme in study 1 (solid dark line) and study 2 (dashed light line) for the coherence weighting of coherentized estimates, as a function of the exponent parameter β in the weighting function (recall we set as $\beta = 15$ in the analyses). In the bottom panel, we see the weighting value $\omega(0.29)$ as a function of β. We note that the minimum of the average BS occurs around $\beta = 20$, which coincides with the point where $\omega(0.29)$ is practically zero. However, there are similar scores when β is between 10 and 40. We also see that the average Brier score from study 1 is less than that of study 2, save the region where β is close to zero. This strongly supports our analysis in §4.4, which shows that independent, coherence-weighted, intraparticipant estimates produce more accurate results.

Figure 7 shows the same analysis as Figure 6, but for the four-way scheme. The top panel shows the average BS of the four-way scheme in study 1 (solid dark line) and study 2 (dashed light line) for the coherence weighting of coherentized estimates as a function of the exponent parameter β in the weighting function. The bottom panel shows the weighting value $\omega(0.32)$ as a function of β. We see the average Brier score of study 1 was less than that of study 2 for a large range.

Figure 6 The Average Brier Score for the Three-Way Coherentization Scheme vs. the Scale Parameter β for Study 1 (Solid Dark Line) and Study 2 (Dashed Light Line) in the Top Panel, and the Weighting Function Evaluated at 0.29 vs. the Scale Parameter β in Bottom Panel

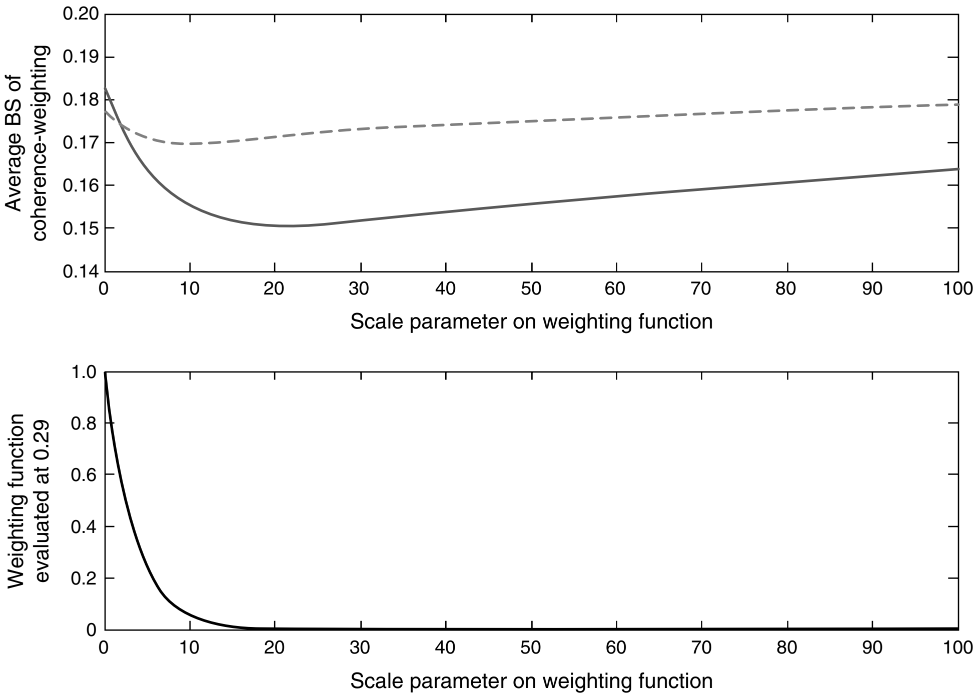
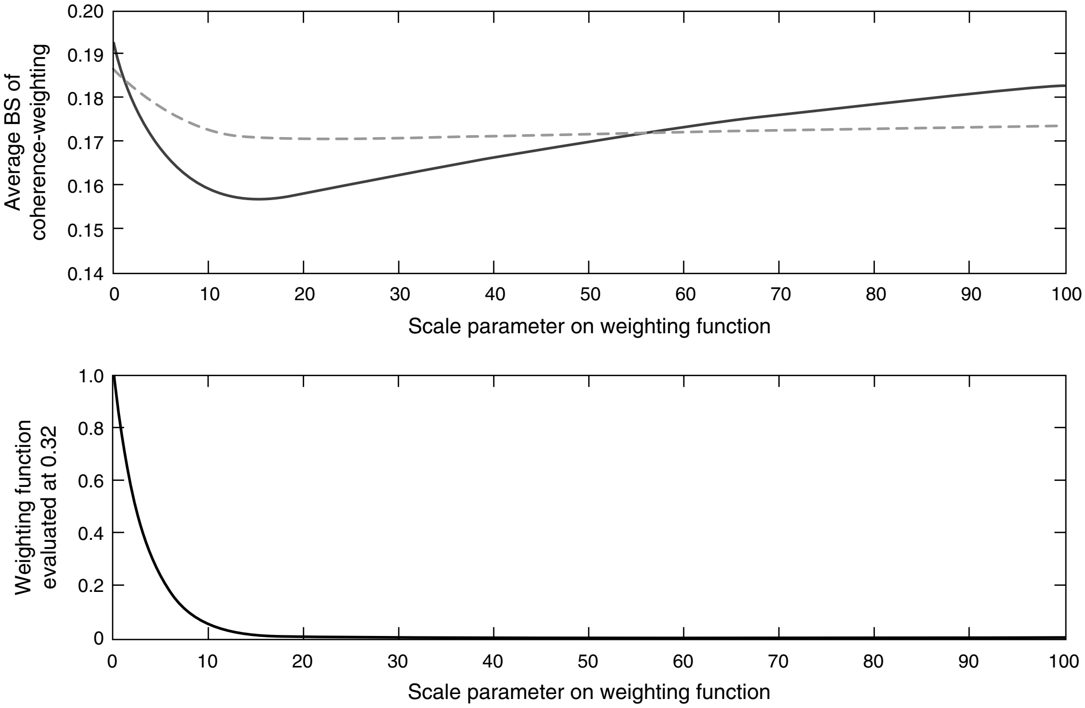


Figure 7 The Average Brier Score for the Four-Way Coherentization Scheme vs. the Scale Parameter β for Study 1 (Solid Dark Line) and Study 2 (Dashed Light Line) in the Top Panel, and the Weighting Function Evaluated at 0.32 vs. the Scale Parameter β in Bottom Panel



Testing different weighting functional forms, we found a linear function from (0, 1) to (0.32, 0) generated similar Brier score results in study 1 for the three-way elicitation (0.1609 with linear versus 0.1515 with original power function), as well as for the four-way scheme (0.1596 with linear versus 0.1568 with original power function), recalling the average Brier score for the equal-weighted averaging of raw judgments was 0.2243.

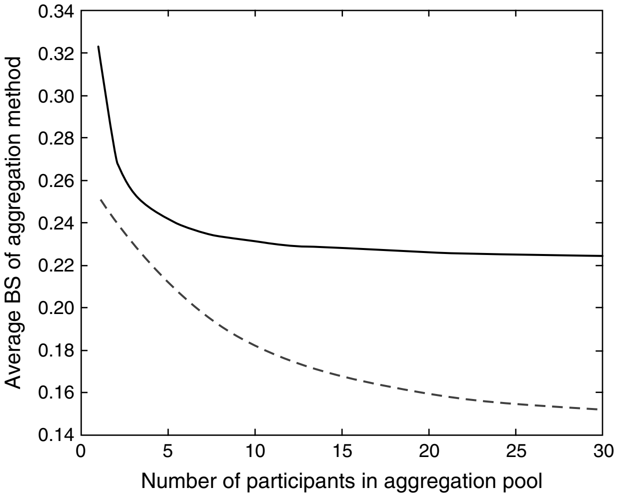
We also found for study 1 that an indicator weighting function that assigned a weight of 1 to the participant for the question if his or her IM was to the left of the critical threshold, and a weight of 0 at or to the right of the critical IM threshold still generated sizable gains for the three-way scheme (0.1726 with indicator function versus 0.1515 with original power function), and the four-way scheme (0.1752 with indicator function versus 0.1568 with original power function). This type of weighting function is similar to the approach of Tsai and Kirlik (2012), yet we allow the number of estimates that are averaged for each question to vary depending on how many judges produce estimates that are to the left of the critical threshold.

For study 2, we found a linear function from (0, 1) to (0.32, 0) generated similar Brier score for the three-way elicitation (0.1698 with linear versus 0.1704 with original power function), as well as for the four-way scheme (0.1747 with linear versus 0.1708 with original power function), recalling the average Brier score for the equal-weighted averaging of raw judgments was 0.2020.

We also found for study 2 that an indicator weighting function that assigned a weight of 1 to the participant for the question if his or her IM was to the left of the critical threshold, and a weight of 0 at or to the right of the critical threshold still generated sizable gains for the three-way scheme (0.1755 with indicator function versus 0.1704 with original power function), and the four-way scheme (0.1812 with indicator function versus 0.1708 with original power function). In sum, these sensitivity results for the weighting function are encouraging for future use of the approach.

Figure 8 shows how the average Brier score results vary depending on the size of the aggregation pool. These results were obtained by randomly sampling

Figure 8 The Average Brier Score of the Equal-Weighted Average (*LINOP*; Solid Line), and the Coherence Weighting of Coherentized Estimates for the Three-Way Coherentization Scheme (Dashed Line), as a Function of the Aggregation Pool Size



1,000 subsets of participants for each integer size ranging from 1 to 30, and then implementing the respective methods and averaging over the samples to get an average overall Brier score. Recalling study 1 had 30 participants, the equal-weighted averaging method (*LINOP*) is shown as the solid black line, and the coherence weighting of coherentized estimates for the three-way scheme is shown in the dashed line, as a function of the pooling size. The initial difference at a pool size of 1 is due to the coherentization of the estimates, and the coherence weighting of coherentized estimates method dominates the *LINOP* for every pooling size.

As the pool size increases, there are diminishing returns in the Brier score improvement for the equal-weighted averaging method, with strong asymptotic returns around a pool size of about 10. Typically, diminishing returns are seen with between three and five experts (Clemen and Winkler 1999, Winkler and Clemen 2004), and this observation is consistent with the equal-weighted averaging method. The coherence weighting of coherentized estimates method however only begins to asymptote around a pool size of 30. This observation suggests that, depending on the degree of incoherence in the pooling population, the decision maker should seek a larger aggregation pool for the coherentization method than for a *LINOP* to increase the quality of the aggregate estimates.

5.3. Generalization of Research

The results of this paper extend past work that measures and uses probabilistic coherence to adjust probability judgments and weight judges in an effort to produce more accurate aggregated forecasts. Probabilistic coherence provides a logical framework to elicit multiple, different subjective probabilities ($P(A)$, $P(B)$, $P(A^c)$, and $P(A \cup B)$), and use these probabilities to adjust the probability of interest ($P(A)$).

For two probability judgments $P(A)$ and $P(A^c)$, creating probabilistic coherence is equivalent to the approach of averaging $P(A)$ and $1 - P(A^c)$, but probabilistic coherence can be applied to any set of variables that are logically related. Probabilistic coherence allows for the most useful information to be elicited. For example, we can ask about a mutually exclusive event B, and the union of two events A and B to make a judgment on A. By comparing the gains in accuracy, we were able to prioritize the best information that should be elicited.

There are key advantages of the approach of this paper when compared to other aggregation approaches. The approach does not require questions to resolve or similar seed questions with known resolution values to be constructed. Using coherence weighting could therefore decrease the time it takes to prepare for the elicitation, when compared with other performance-based weighting schemes that use seed variables. For example, there is a significant time burden in constructing the seed variables necessary for Cooke's classical weighting method (Cooke and Goossens 2008), especially when the number of seed variables needed is large (Clemen 2008). The seed variables also need to closely match the theme of the real forecasting events.

With the three-way scheme, only one extra event B is constructed, and $A \cup B$ then follows. The event B *directly* concerns the target event A. We found that constructing B is usually straightforward, and the three-way coherentization scheme would be easily translated to real forecasting questions. For example, given an upcoming election, the events A and B could describe the various candidates. Alternatively, for forecasting growth of gross domestic product, the events A and B could be different intervals of growth.

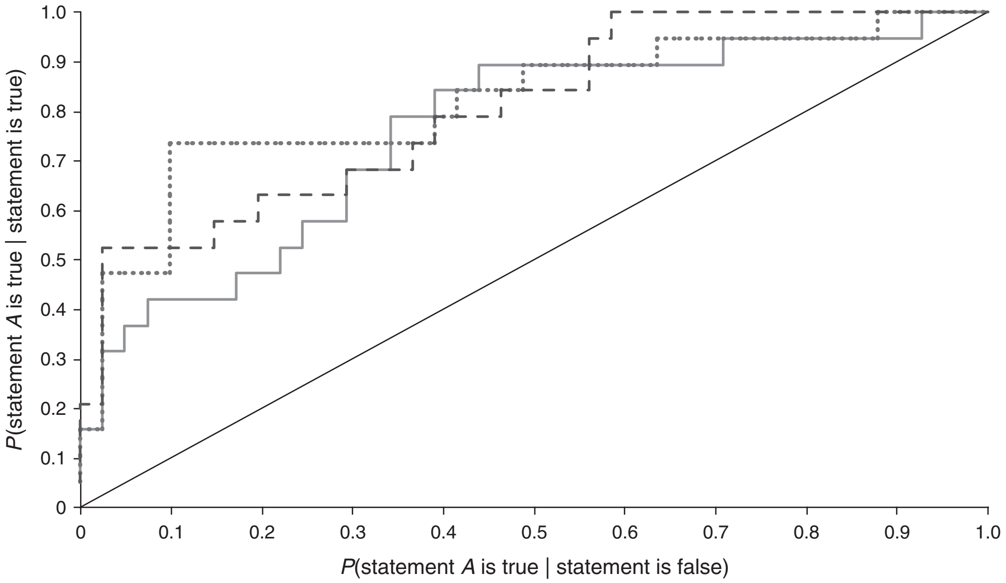
We found that eliciting $P(A^c)$ does not generate large accuracy gains for $P(A)$ when compared to eliciting $P(B)$ and $P(A \cup B)$. There are two potential reasons for this. First, for many questions, there was no convenient way to express A^c, other than to simply say "not A." Thus, participants might have been *anchoring* on $P(A)$ when providing $P(A^c)$, even if the judgments were elicited independently (Tversky and Kahneman 1974). Eliciting $P(B)$ perhaps allowed the participant to think about $P(B)$ in a manner that did not anchor on $P(A)$. Second, epistemic uncertainty or ignorance can pass as a coherent estimate, and it is not possible to establish a critical IM threshold for answering 0.5 for the probabilities $P(A)$ and $P(A^c)$ with the two-way coherentization scheme as can be done with the three- and four-way coherentization schemes (IM values of 0.29 and 0.32, respectively).

Coherentizing related probability estimates for the three- and four-way schemes always increased the accuracy of the aggregated individual estimate, thus providing support for "crowdsourcing" *within* an individual (Herzog and Hertwig 2009, Vul and Pashler 2008, Larrick and Soll 2006). Moreover, in line with previous research (Hirt and Markman 1995, Lord et al. 1984, Mandel 2005, Sieck et al. 2007, Williams and Mandel 2007), we found that participants' accuracy and coherence were improved by judging logically-related events in a concurrent as opposed to spaced manner. Thus, concurrent judgments may represent a preferable elicitation mode for improving judgment quality in contexts where judgments will be used without further aggregation or transformation. However, with coherence weighting, we found increasing the independence of the related judgments was more effective in improving the accuracy of the aggregate judgments than eliciting them concurrently. Our best methods produced gains that were over 30% better than the *LINOP*, which is in line with the gains seen by others (Tsai and Kirlik 2012, Wang et al. 2011). Understanding this elicitation-aggregation trade-off can be important for reaching decisions about the optimal means for leveraging forecasts or other advice that comes in the form of probabilistic judgments.

In terms of eliminating the fifty–fifty blip (Bruine de Bruin et al. 2002), the current approach effectively removed the 0.5 probabilities that should likely not be interpreted as point estimates, but rather that represent epistemic uncertainty. We were also able to justify the reduced influence of probabilities other than 0.5 by using the incoherence metric. Reducing the number of 0.5 estimates allowed better *discrimination*, which is shown in Figure 9 in the form of ROC curves of the equal-weighted averages of the raw judgments (solid, light gray line), and the three-way (dashed, dark gray line) and four-way (dotted, gray line) coherentization schemes. Whereas we do not see complete dominance over the equal-weighted average of raw judgments, we do see a significant advantage of the two coherence-weighting methods where the response probabilities are between 0.4 and 0.6. (This region is not immediately discernible from Figure 9, but corresponds approximately to the region between 0.1 and 0.3 on the x axis, and 0.3 and 0.8 on the y axis.)

We also observe better discrimination and calibration when we decompose the Brier score for the three- and four-way schemes. We do so for study 1, where we observed the largest performance improvement after coherence weighting. Using a three-part decomposition of the Brier score ($\text{BS} = \text{uncertainty} - \text{discrimination} + \text{calibration}$) (Yaniv et al. 1991, Murphy 1973) with six equally spaced subjective probability partitions, we have for the *LINOP* in study 1, $\text{BS} = 0.223 = 0.216 - 0.058 + 0.064$. For the three-way scheme, we have $\text{BS} = 0.150 = 0.216 - 0.082 + 0.016$, and we have $\text{BS} = 0.152 = 0.216 - 0.081 + 0.017$ for the four-way scheme. Discrimination over uncertainty, η^2, captures the proportion of variance explained in the outcomes by the judgment categories (Sharp et al. 1988). In study 1, $\eta^2 = 0.269$ for the *LINOP*, $\eta^2 = 0.375$ for the three-way scheme, and $\eta^2 = 0.380$ for the four-way scheme. Thus, either coherence-weighting scheme yielded slightly more than a 40% increase over the LINOP in explaining outcome variance—a substantial proportional increase and over a 10% increase in explained variance in absolute terms. In terms of calibration, it is useful to consider the square root of the calibration index taken from the Brier decomposition since it represents the average absolute deviation from perfect calibration. In study 1, the square roots are 0.253 for the *LINOP*, 0.126 for the three-way scheme, and 0.130

Figure 9 The ROC Curves for Study 1 for the Equal-Weighted Averaging of Raw Judgments (*LINOP*, Solid, Light Gray Line), the Coherence Weighting of Coherentized Estimates for the Three-Way Coherentization Scheme (Dashed, Dark Gray Line), and the Coherence Weighting of Coherentized Estimates for the Four-Way Coherentization Scheme (Dotted Gray Line)



for the four-way scheme. Thus, we see that the proportional increase in calibration close to 50%, an even more substantial increase than we observed for discrimination. Thus, both calibration and discrimination are substantially improved by coherence weighting when compared with the *LINOP*.

Within the spectrum of technical complexity, the coherence-based weighting approach of this paper is much simpler than the approach of Wang et al. (2011), and comparable to that of simple, equal-weighted averaging. The coherentization algorithm is done for each participant-question pair, and thus the procedure is linear in each factor. The entire analysis can be closely approximated in spreadsheet software for the three-way scheme, without the need of sophisticated computational algorithms. We found the degree to which a set of judgments is incoherent can be approximated as,

$$\sqrt{(P(A)+P(B)-P(A\cup B))^2}$$

and the coherentizing can be closely approximated by distributing this difference $P(A)+P(B)-P(A\cup B)$ equally among the three probabilities.

In general, the findings of this research are applicable in any situation where expert probability forecasts are aggregated. In particular, the findings could be used for aggregating the responses of multiple experts for use within a shared model. For example, in the Bayesian network case model of Karvetski et al. (2013), each conditional probability distribution for an arc requires elicitations over two or three mutually exclusive and exhaustive states in a probability space (that match with A, B, and $A\cup B$), and, in total there are 115 probability judgments that are needed. The judgments could be elicited in a manner similar to study 1.

5.4. Future Work

We recognize that our findings are contingent on the three performance measures that were used (average Brier score, number of questions one method improved on another, and average absolution distance of improvement), and that if additional performance measures (e.g., averaged log score, correlation, slope) were used, they could change the degree (although, unlikely, the direction) of preference among the methods.

One factor that we explicitly changed across the two studies was the independence of the intraparticipant judgments. For the first study, we spaced the related judgments out across the 60 questions, and for the second study, the related judgments were elicited consecutively. However, if a decision maker seeks a forecast for one key event, it is not feasible to generate 59 other statements to get independent estimates. Future work might address how far apart to temporally space the related subjective probability elicitations to realize optimal forecasting gains, and how much of a decrement in gain accrues as the window size is reduced. This type of analysis might parallel the type undertaken in studying optimal spacing between test sessions to enhance learning (Rohrer and Pashler 2007).

For some risk and decision analyses, experts would have at least some training in probability biases as well as formal elicitation methods, and best practice would include having an analyst interview them and work with each expert one-on-one. Many expert settings are clearly different from the circumstances under which the undergraduates in our studies provided their probabilities. Future research could look at how our results apply to real forecasting questions with real experts, such as pundits' forecasts of election outcomes and other key events that face policymakers. Part of this research would investigate if the gains observed within our studies would still be achievable, and, if so, why. For example, in our studies, we were not incentivizing performance. Some participants likely took the survey more seriously than others, and it is possible that variation in incoherence was associated with experimental vigilance by the participant. Future work could investigate this potential relationship by screening for different levels of vigilance. For instance, Oppenheimer et al. (2009) developed a simple one-response task aimed at detecting whether a participant is an experimental satisficer (namely, one who does not follow the task instructions properly because of a purported lack of cognitive effort devoted to the task). This task could be used to create low- and high-vigilance groups in future research. As well, even with experts, future work could investigate the utility of the approach when time is limited and working memory is accordingly taxed (Sprenger et al. 2011).

Future work could continue to investigate how many forecasters and what degree of coherence are needed to efficiently generate gains. There will likely be diminishing performance gains as these values increase (Winkler and Clemen 2004). Alternatively, future work could investigate the calibration of the coherence-weighted estimates, in an effort to better understand when to push the aggregative estimates outside of the convex combination of forecast values, or when to weight forecasters with negative weights. Finally, we suggest comparing coherence weighting with Cooke's classical weighting method (Cooke 1991), both in terms of accuracy and ease of implementation. This would allow us to examine the relative utility of coherence weighting for forecasts of continuous variables.

Acknowledgments

The authors acknowledge the helpful comments from researchers within the IARPA ACE program. The U.S. and Canadian governments are authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the Canadian or U.S. governments. This research was supported in part by the IC Postdoctoral Research Fellowship Program, by the Natural Sciences and Engineering Research Council of Canada Discovery Grant [249537-2007], and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center [Contract D11PC20062].

Appendix A. Additional Statements Evaluated by Participants

Additional examples of A statements used in the two studies with the truth values in parentheses (complete list available from corresponding author).

- In the Earth's solar system, Mars is the fifth plant from the sun (F).
- Michelangelo painted the Sistine Chapel (T).
- Hydrogen is the first element listed in the periodic table (T).
- As of 2008, Nebraska is the top corn-producing state in the United States (F).
- In terms of 2011 population, Manhattan is the largest of the five New York City Boroughs (F).
- Volvo is a Swedish car manufacturer (T).
- Massachusetts was the first state admitted to the United States (F).
- The Pacific Ocean is the largest of Earth's oceans (T).
- The United States won the most total medals in the 2008 Beijing Olympics (T).
- Richard Nixon was the 37th president of the United States (T).
- The average annual rainfall in Seattle is between 30 and 40 inches (T).
- Melbourne is the capital of Australia (F).

Appendix B. Comparing Numeracy and Coherence Within Study 2

In study 2, we examined the role of numeracy in the accuracy and coherence of participants' judgments. Numeracy is the ability of people to use numeric information and to reason with numerical concepts, and it has been shown to vary greatly across individuals (Peters et al. 2007). We were unsure whether more numerate people would make more coherent or more accurate estimates for the general knowledge statements we used. The most well-developed test of numeracy is a series of mathematical problems varying in their difficulty (Weller et al. 2012). This test produces a roughly normal distribution of scores among a general population by primarily asking questions about probabilities, and we chose to use it for measuring participants' numeracy after they completed all other survey items in study 2.

Out of eight questions, the average numeracy score was 4.32 in study 2. Scores ranged from one to seven among participants. Participants' probability estimates y_1^i were weighted by their numeracy similarly to how they had otherwise been weighted by their coherence, but here all estimates from a participant received the same weight. The highest numeracy score received the highest weight. Decreasing scores received decreasing weights provided by a power function for which the best parameter value was 6.16. Numeracy weighting of individuals' responses produced an improvement in the Brier score of 3.131% over the equal-weighted average of raw estimates (0.1957 versus 0.2020). The proportion of questions for which accuracy improved (34/60) was not statistically significant ($t_{59} = 1.042$, $p = 0.151$). The average absolute value of improvement also did not reach statistical significance ($t_{59} = 1.648$, $p = 0.052$).

The small increase in accuracy from numeracy weighting of participants compared to coherence weighting is not surprising when we consider the weak correlation between the incoherence metric and numeracy scores ($r = -0.147$). Perhaps because the questions on the numeracy scale are worded as math problems and our questions are not, numeracy did not relate to how well people followed rules of probability for responding to general knowledge statements in study 2. Whatever the reason, numeracy scores did not significantly correlate with participants' incoherence of estimates ($t_{26} = -0.758, p = 0.226$).

References

Brier GW (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Rev.* 78(1):1–3.

Bruine de Bruin W, Fischbeck PS, Stiber NA, Fischhoff B (2002) What number is "fifty–fifty"?: Redistributing excessive 50% responses in elicited probabilities. *Risk Anal.* 22(4):713–723.

Burgman MA, McBride M, Ashton R, Speirs-Bridge A, Flander L, Wintle B, Fidler F, Rumpff L, Twardy C (2011) Expert status and performance. *PLoS One* 6(7):1–7.

Clemen RT (2008) Comment on Cooke's classical method. *Reliability Engrg. System Safety* 93:760–765.

Clemen RT, Winkler RL (1999) Combining probability distributions from experts in risk analysis. *Risk Anal.* 19(2):187–203.

Cooke RM (1991) *Experts in Uncertainty* (Oxford University Press, Oxford, UK).

Cooke RM, Goossens LLHJ (2008) TU Delft expert judgment data base. *Reliability Engrg. System Safety* 93(5):657–674.

De Finetti B (1990) *Theory of Probability: A Critical Introductory Treatment* (John Wiley & Sons, New York).

Genest C, McConway KJ (1990) Allocating the weights in the linear opinion pool. *J. Forecasting* 9(1):53–73.

Gneiting T (2011) Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* 106(494):746–762.

Herzog SM, Hertwig R (2009) The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psych. Sci.* 20(2):231–237.

Hirt ER, Markman KD (1995) Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *J. Personality Soc. Psych.* 69(6):1069–1086.

Hsee C (1996) The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organ. Behav. Human Decision Processes* 67(1): 247–257.

Karvetski CW, Olson KC, Gantz DT, Cross GA (2013) Structuring and analyzing competing hypotheses with Bayesian networks for intelligence analysis. *EURO J. Decision Processes*, ePub ahead of print April 30, http://link.springer.com/article/10.1007/s40070-013-0001-x.

Kolmogorov A (1956) *Foundations of the Theory of Probability* (Chelsea Publishing Company, New York).

Larrick RP, Soll JB (2006) Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Sci.* 52(1):111–127.

Lindley DV, Tversky A, Brown RV (1979) On the reconciliation of probability assessments. *J. Roy. Statist. Soc.* 142(2):146–180.

Lord CG, Lepper MR, Preston E (1984) Considering the opposite: A corrective strategy for social judgment. *J. Personality Soc. Psych.* 47(6):1231–1243.

Macchi L, Osherson D, Krantz DH (1999) A note on superadditive probability judgment. *Psych. Rev.* 106(1):210–214.

Mandel DR (2005) Are risk assessments of a terrorist attack coherent? *J. Experiment. Psych.: Appl.* 11(4):277–288.

Mandel DR (2008) Violations of coherence in subjective probability: A representational and assessment process account. *Cognition* 106(1):130–156.

Merrick JRW (2008) Getting the right mix of experts. *Decision Anal.* 5(1):43–52.

Murphy AH (1973) A new vector partition of the probability score. *J. Appl. Meteorology* 12(4):595–600.

Oppenheimer DM, Meyvis T, Davidenko N (2009) Instructional manipulation checks: Detecting satisficing to increase statistical power. *J. Experiment. Soc. Psych.* 45(4):867–872.

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

Pate-Cornell ME (1996) Uncertainties in risk analysis: Six levels of treatment. *Reliability Engrg. Systems Safety* 54(2–3):95–111.

Peters E, Dieckmann NF, Dixon A, Hibbard JH, Mertz CK (2007) Less is more in presenting quality information to consumers. *Medical Care Res. Rev.* 64(2):169–190.

Predd JB, Osherson DN, Kulkarni SR, Poor HV (2008) Aggregating probabilistic forecasts from incoherent and abstaining experts. *Decision Anal.* 5(4):177–189.

Rohrer D, Pashler H (2007) Increasing retention without increasing study time. *Current Directions Psych. Sci.* 16(4):183–186.

Sharp GL, Cutler BL, Penrod SD (1988) Performance feedback improves the resolution of confidence judgments. *Organ. Behav. Human Decision Processes* 42(3):271–283.

Sieck WR, Merkle EC, Van Zandt T (2007) Option fixation: A cognitive contributor to overconfidence. *Organ. Behav. Human Decision Processes* 103(1):68–83.

Sprenger AM, Dougherty MR, Atkins SM, Franco-Watkins AM, Thomas RP, Lange N, Abbs B (2011) Implications of cognitive load for hypothesis generation and probability judgment. *Frontiers Psych.* 2(129):1–15.

Surowiecki J (2005) *The Wisdom of Crowds* (Double Day, New York).

Tetlock PE (2005) *Expert Political Judgment: How Good Is It? How Can We Know?* (Princeton University Press, Princeton, NJ).

Tsai J, Kirlik A (2012) Coherence and correspondence competence: Implications for elicitation and aggregation of probabilistic forecasts of world events. *Proc. Human Factors and Ergonomics Soc. 56th Annual Meeting* (Sage, Thousand Oaks, CA), 313–317.

Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185(4157):1124–1131.

Tversky A, Koehler DJ (1994) Support theory: A nonextensional representation of subjective probability. *Psych. Rev.* 101(4):547–567.

Vul E, Pashler H (2008) Measuring the crowd within: Probabilistic representations within individuals. *Psych. Sci.* 19(7):645–647.

Wang G, Kulkarni SR, Poor HV, Osherson DN (2011) Aggregating large sets of probabilistic forecasts by weighted coherent adjustment. *Decision Anal.* 8(2):128–144.

Weller JA, Dieckmann NF, Tusler M, Mertz CK, Burns WJ, Peters E (2012) Development and testing of an abbreviated numeracy scale: A Rasch analysis approach. *J. Behavioral Decision Making* 26(2):198–212.

Williams JJ, Mandel DR (2007) Do evaluation frames improve the quality of conditional probability judgment? McNamara DS, Trafton JG, eds. *Proc. 29th Annual Meeting of the Cognitive Sci. Soc.* (Lawrence Erlbaum Associates, Inc., Mahwah, NJ), 1653–1658.

Winkler RL, Clemen RT (2004) Multiple experts vs. multiple methods: Combining correlation assessments. *Decision Anal.* 1(3):167–176.

Wright G, Rowe G, Bolger F, Gammack J (1994) Coherence, calibration, and expertise in judgmental probability forecasting. *Organ. Behav. Human Decision Processes* 57(1):1–25.

Yaniv I, Yates JF, Smith JEK (1991) Measures of discrimination skill in probabilistic judgment. *Psych. Bull.* 110(3):611–617.

Christopher W. Karvetski is a quantitative finance analyst within the Model Validation and Analytics Group, Financial Intelligence Unit at Bank of America, where his focus is on developing models for detecting and mitigating financial crimes risk, including money laundering, terrorist financing, and economic sanctions risk. Previously, he was an Intelligence Community Research Fellow and assistant professor in the Applied Information Technology Department at George Mason University in Fairfax, Virginia. His main interests are in risk and decision analysis, and predictive modeling and analytics, with a focus on defense and security applications. He holds a Ph.D. in systems engineering from the University of Virginia. He is a member of the Society for Risk Analysis, the Decision Analysis Society, and Tau Beta Pi.

Kenneth C. Olson is an assistant professor in the Volgenau School of Engineering at George Mason University. He earned his Ph.D. in quantitative psychology from The Ohio State University. He completed a fellowship with the Intelligence Community Postdoctoral Research Fellowship Program. He is most knowledgeable about social, cognitive, and motor processes in decision making and has extensive experience educating both the public and professionals on common judgment errors. His teaching and research provide statistical modeling tools to non-statisticians, including work to develop a structured analytic technique to implement Bayesian networks for intelligence assessments. In recent years, he has collaborated on several projects to improve international political forecasting. He is a member of the Society for Mathematical Psychology, the Cognitive Science Society, the Society for Risk Analysis, and the Society for Judgment and Decision Making.

David R. Mandel is a senior scientist in the Sensemaking and Decision Group of the Socio-Cognitive Systems Section at DRDC Toronto, which is part of the Government of Canada's Department of National Defence. He is also adjunct professor of psychology at York University. He holds a Ph.D. in psychology from the University of British Columbia. His research focuses on basic and applied topics in judgment and decision making, with particular emphasis on the application of such work to the defense and security sector. He has served as a scientific advisor to such organizations as the The National Academies, The National Institutes of Health, The Office of the Director of National Intelligence, the U.S. Department of Defense, and NATO. His books include *The Psychology of Counterfactual Thinking* (Routledge 2005) and *Neuroscience of Decision Making* (Psychology Press 2011).

Charles R. Twardy is a research assistant professor at George Mason University where he leads the SciCast forecasting project, a four-year effort to enhance the accuracy, calibration, and timeliness of crowdsourced forecasts of (previously) geopolitical events and (now) science and technology. His work focuses on fusing Bayesian networks with prediction markets for crowdsourced elicitation of large joint probability spaces. More generally he is interested in inference and decision making with a special interest in causal models. Previous work includes counter-IED models, credibility models, sensor selection, hierarchical fusion, and epidemiological models. He holds a dual Ph.D. in history and philosophy of science and cognitive science from Indiana University.