Supplementary Information for Structured Expert Judgment

Roger Cooke
June 22, 2022

Online sources:
website for Structured Expert Judgment http://www.cooke-aspinall.net/
Wiki page: https://en.wikipedia.org/wiki/Structured_expert_judgment:_the_classical_model
Selected publications

- Int'l J of Forecasting SEJ what do the Data Say 2021
- CDC Burden of Disease Food & Water pathways 2021
- PNAS SEJ ice sheets and SI SEJ data and elicitation protocol 2019
- REEP Validation for Classical Model 2018
- RESS Cross Validation extensive SOM on validation, aggregation 2017
- Comp & OR Quantifying Info Security Risks 2012
- PLoS 1 Evaluation of Performance: WHO SEJ study of global burden of disease
- Elementa Stormwater Management in Chesapeake Bay; extensive SI elicitation protocol 2015
- Cons Bio Impacts of Asian Carp Invasion Lake Erie - SEJ SI 2015
- RFF SEJ Breast Feeding and IQ: AStA 2019 Risk Analysis 2021
- RFF SEJ IQ and Earnings
- Nature Climate Change: Messaging Climate Uncertainty with extensive SI on representation of uncertainty 2014

Videos

- Intro to SEJ (10 mins)
- The Confidence Trap (10 mins)
- Ice sheets (11 mins)
- Validation (25 mins)

Online Course
SEJ TU Delft

Contents

1. Technical Details of the Classical Model

The classical model for evaluating and combining experts considers experts as statistical hypotheses and prioritizes statistical accuracy. Informativeness is also important and serves to modulate between experts with similar statistical accuracy so that it is impossible to compensate poor statistical accuracy with very high information. Performance weights for combining experts are derived which satisfy asymptotic strictly proper scoring rule constraints.

This exposition considers five assessed quantiles (i.e., *5%*, *25%*, *50%*, *75%* and *95%*) for each elicited uncertain quantity, or item. The expert could be a human assessor, a computer code, or some combination of the two.

1.1. Statistical Accuracy (aka Calibration[i])

[i] In the science/engineering world, "calibration" denotes removing bias in measurement instruments. The psychology community introduced the notion of scoring experts' statistical accuracy and termed the result "calibration". As this causes

The assessed quantiles divide the range of possible values into six inter-quantile intervals for which an expert's probabilities are known (e.g., $p_1 = 0.05$ [less than or equal to the *5%* quantile]; $p_2 = 0.20$ [greater than the *5%* quantile and less than or equal to the *25%* quantile], etc.). If *N* quantities are assessed as calibration questions, or items, each expert may be regarded as a statistical hypothesis, namely that each realization falls in one of the six inter-quantile intervals with the probability vector:

$$p = (0.05, 0.20, 0.25, 0.25, 0.20, 0.05)$$

Suppose we have realizations $x_1, \ldots, x_N$ of these quantities. We may then form the sample distribution of the expert's inter-quantile intervals as $s_e = (s_{1,e}, s_{2,e}, \ldots, s_{6,e})$, where:

$$s_{1,e} = \#\{i \mid x_i \leq 5\% \text{ quantile}\} / N$$

$$s_{2,e} = \#\{i \mid 5\% \text{ quantile} < x_i \leq 20\% \text{ quantile}\} / N$$

$$\ldots$$

$$s_{6,e} = \#\{i \mid 95\% \text{ quantile} < x_i\} / N$$

Note that the sample distribution depends on the assessments from the expert, *e*. If the realizations are indeed drawn independently from a distribution with quantiles as stated by the expert, then the quantity:

$$2N \sum_{i=1}^{6} s_{i,e} \ln\left(\frac{s_{i,e}}{p_i}\right) \tag{1.1}$$

is asymptotically distributed as a chi-square random variable with five degrees of freedom. This is the likelihood ratio statistic. If we extract the leading term of the logarithm, we obtain the familiar chi-square test statistic for goodness of fit. If, after a few realizations, the expert saw that all realizations fell outside of their *90%* central confidence interval, (s)he might conclude that their intervals were too narrow and broaden them on subsequent assessments. This means that, for this expert, the uncertainty distributions are *not* independent, and (s)he learns from the realizations. Expert learning is not a goal of a study, and their joint distribution is not elicited. Rather, the analyst wants experts who do not need to learn from the elicitation. Hence the analyst scores the expert, *e*, as the statistical likelihood of the hypothesis:

H_e: The inter-quantile interval containing the true value for each item is drawn independently from probability vector, p.

A simple test for this hypothesis uses the test statistic (1.1), and the likelihood, or p-value, or statistical accuracy score (aka 'calibration score') of this hypothesis, is:

$$C_e = P(\chi \geq r \mid H_e) \tag{1.2}$$

where χ is a chi-square distributed random variable with five degrees of freedom and *r* is the value of (1.1) based on the observed values $x_1, \ldots, x_N$. It is the probability under hypothesis H_e that a deviation at least as great as *r* should be observed on *N* realizations if H_e were true. Calibration scores are absolute and can be compared across studies. However, before doing so, it is appropriate to equalize the power of the different hypothesis tests by equalizing the effective number of realizations. To compare scores on two data sets with *N* and *N'* realizations, we simply use the minimum of *N* and *N*' in (1.1), without changing the sample distribution, *s*.

Although the calibration score employs the language of simple hypothesis testing, it must be emphasized that we are not rejecting expert-hypotheses; rather we are using this language to measure the degree to which the data supports the hypothesis that the expert's probabilities are statistically accurate. Low scores, near zero, mean that it is unlikely that the expert's probabilities are correct. A common thumb rule in testing a simple multinomial hypothesis is that there should be at least *5* expected observations in each cell. In practice this is often relaxed to one expected observation. With *10* calibration variables, there is *½* expected observation in the lowest and highest cells. Simulation studies [Cooke 2014] show that the statistical power afforded by *10* observations enables only the identification of gross differences in statistical performance. The default choice

insuperable confusion when explaining structured expert judgment to scientists and engineers, "calibration" is replaced "statistical accuracy" despite its currency in the psychological community.

of *10* calibration variables reflects a compromise allowing coarse distinctions without excessively burdening the elicitation. This compromise is motivated by the fact that large differences are usually present in expert panels [Colson and Cooke 2017], bearing in mind that CM is not testing statistical hypotheses as such.

1.2. Information

Measuring information requires associating a density to each assessment for each item. To do this, we use the unique density that complies with the expert's quantiles and is minimally informative with respect to a background measure. For a uniform background measure, the density is constant between the assessed quantiles, and is such that the total mass between the quantiles agrees with *p*. The background measure is not elicited from experts as it must be the same for all experts; instead, it is chosen by the analyst. The Classical Model (CM) uses by default either the uniform or the log-uniform background measure, as these have no location parameter other than the support of the background measure.

The uniform and log-uniform background measures require an 'intrinsic range' on which these measures are concentrated. The CM implements the so-called '*k% overshoot rule*'. For each item, we consider the smallest interval, $I = [L, U]$, containing all the assessed quantiles of all experts and the realization, if known. This interval is extended to:

$$I^* = [L^*, U^*]; L^* = L - k\,(U - L) / 100; U^* = U + k\,(U - L) / 100 \qquad (1.3)$$

The value of *k* is chosen by the analyst. A large value of *k* tends to make all experts look quite informative and tends to suppress the relative differences in information scores. The information score of the expert, *e*, on assessments for uncertain quantities *1, …, N,* is:

$$Inf_e = (1 / N) \sum_{i = 1, \ldots, N} I\,(f_{e,\,i} \mid g_i) \qquad (1.4)$$

where g_i is the background density for item *i* and $f_{e,\,i}$ is the fitted density for expert, *e*, for item *i*. This is proportional to the relative information of the expert's joint distribution given the background, under the assumption that the items are independent. As with calibration, the assumption of independence reflects a desideratum of the analyst and is not an elicited feature of the expert's joint distribution. The information score does not depend on the realizations. An expert can acquire a high information score by choosing their quantiles very close together.

Evidently, the information score of an expert depends on the intrinsic range and on the assessments of the other experts. Hence, information scores cannot be rigorously compared across studies.

Both statistical accuracy and informativeness are dimensionless. The information score is a 'slow' function, that is, large changes in the assessments produce only modest changes in the information score. Information scores in a panel of experts typically vary by a factor less than *3* whereas the statistical accuracy (1.1) varies over several orders of magnitude. This ensures that the normalized product of statistical accuracy and information is driven by the calibration score. It also means that modest changes in informativeness correspond to sizeable changes in the distributions. Increasing informativeness by a factor of two roughly corresponds to halving the distance between the 95^{th} and 5^{th} percentiles.

1.3. Scoring Rules for Individual Items

Scoring rules were introduced by de Finetti in 1937 as tools for encouraging honesty in eliciting subjective probabilities and have been further developed by many authors (Shuford and Massengill 1966). An expert receives a score as a function of their probability assessment and the realization. The score is strictly proper if the expert maximizes the expected score per item by, and only by, stating their true belief. Using a result of Murphy (1977), DeGroot and Fienberg (1983) gave an additive decomposition of strictly proper rules into 'calibration' and 'refinement' terms, thereby replacing Murphy's 'resolution' (refinement applies only to well-calibrated experts). In the case of the logarithmic rule, refinement becomes Kullback-Leibler Divergence (aka, directed divergence, cross entropy, or relative information [the term we use]).

Scoring rules for individual variables were not designed for evaluating or combining experts and are not generally fit for that purpose. Indeed, rewarding honesty is not the same as rewarding quality. A simple example explains this difference: Consider *100*-coin tosses. An expert assesses the probability of heads on each toss as ½. The score for the outcome heads is the same as their score for tails on each toss. If the score

for all *100* assessments is a function of their *100* scores for the individual tosses, then their score for *100* tosses is independent of the outcome sequence; the outcome of *100* heads receives the same score as *50* heads and *50* tails.

Another example concerns the quadratic rule for '*rain / no rain*'. This rule is positively sensed on *[-1.1]* and assigns the score $2r - r^2 - (1-r)^2$ if rain occurs and *r* is the expert's probability of rain. Interchange *r* and *(1– r)* in case it doesn't rain. Consider *1000* next day forecasts of rain by two experts. The experts bin their forecasts as shown below:

Table SI-1

Probability of Rain next day:		5%	15%	25%	35%	45%	55%	65%	75%	85%	95%	Totals
expert 1	assessed	100	100	100	100	100	100	100	100	100	100	1000
	realized	5	15	25	35	45	55	65	75	85	95	500
expert 2	assessed	100	100	100	100	100	100	100	100	100	100	1000
	realized	0	0	0	0	0	100	100	100	100	100	500
average quadratic score	expert 1	0.67										
	expert 2	0.84										

The experts are equally informative in the sense that they assign the same probabilities to the same number of days. *Expert 1* is statistically perfectly accurate whereas *expert 2* is massively inaccurate. Woe unto anyone basing his decisions on *expert 2's* forecasts. Nonetheless, *expert 2's* quadratic score is higher than that of *expert 1*. For more discussion see (Cooke 1991, 2014, Colson and Cooke 2017).

The Probability Interval Score *(PIS)* and its related Continuous Ranked Probability Scores (*CRPS)* have recently been applied to COVID-19 probabilistic predictions (Ray et al 2020), so it is appropriate to include a brief discussion of these. Numerical insight into these scores requires a bit more effort, so more detail is provided.

For the $(1-\alpha)$ interval *[L, U]* with upper (lower) bound *U (L)*, the *PIS* (negatively sensed) for realization *y* is $(U - L) + (2 / \alpha) \times [(L - y)_+ + (y - U)_+]$ where $X_+ = X$ if $X > 0$ and $= 0$ otherwise. $s = 2 / \alpha$ is the slope of the overconfidence penalty for $Y \notin [L, U]$. The length *(U - L)* is called the 'sharpness'; small values reward concentrated probability mass. If $Y \sim Unif[0, 1]$, the central *0.9* interval is *[0.05, 0.95]* with expected *PIS*:

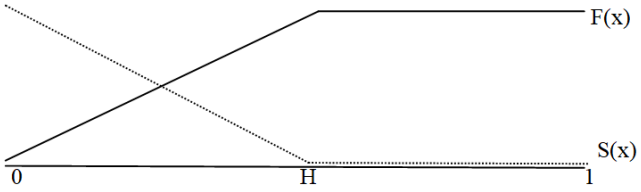
$$0.9 + 2 \times \int_{0 \ldots 0.05} s \times u \, du = 0.9 + (2 / 0.1) \times 0.05^2 = 0.95$$

The integral is doubled to account for $Y > U$. Suppose an expert prefers to give an *80%* interval *[0.1, 0.9]*, $s = 2 / 0.2 = 10$. The expected score is $0.8 + 2 \times \int_{0 \ldots 0.1} s \times u \, du = 20 \times ½ \times 0.1^2 = 0.9 < 0.95$. An expert seeking to optimize (i.e., minimize) their expected score might take a central *2%* prediction interval *[0.49, 0.51]* with expected score $0.02 + 2 \times 2 / 0.98 \times 0.49^2 / 2 = 0.51$ (or take $lim_{\epsilon \to 0} [,5 - \epsilon, 5 + \epsilon]$ with expected score ½). All these prediction intervals have zero information relative to the uniform background measure on *[0, 1]*, so from that viewpoint there isn't much to choose.

The way in which the *PIS* trades overconfidence for sharpness may strike some as counter intuitive. For example, an expert claiming the degenerate interval *[0.5, 0.5]* has *40%* probability of catching the realization would achieve an expected score of *0.833*, better than the score of the central *90%* interval. The sharpness of an interval of zero length outweighs the overconfidence of claiming *40%* mass at the point *0.5*. Of course, this example is blocked if probability intervals are required to be *90%*; assigning *90%* mass to the point *0.5* returns an interval score of *5*. Such scores from several experts could cause bad statistical performance, depending on how the experts are combined.

The weighted *PIS* converges to the continuous ranked probability score (*CRPS*). Applied to a set of probabilistic forecasts and realizations, the expected *CRPS* is based not on what the expert believes, but on the realizations. For illustration we assume that realizations are sampled from variable *Y* uniformly distributed on *[0, 1]*.

Consider an assessment of *Y* by an expert whose distribution is $X \sim Unif[0, H]$, $H \leq 1$. The expert thinks values above *H* are impossible, although these can in fact arise. The *CDF* of *X*, $F(x) = (x / H)$ and the survivor function of *X*, $S(x) = (1 - CDF$; dotted) are pictured below:



The expected *CRPS* is: $\int_{y=0\ldots1}\int_{x=0\ldots1}(F(x)-1_{\{x\geq y\}})^2\,dx\,dy$. The calculation is broken into 2 steps:

$y \leq H$:

$$\int_{y=0\ldots H}\left[\int_{x=0\ldots y}(x/H)^2\,dx+\int_{x=y\ldots H}((H-x)/H)^2\,dx\right]dy$$

$$=\int_{y=0\ldots H}\left[y^3/(3H^2)+\int_{z=H-y\ldots0}z^2/H^2\,(-dz)\right]dy=$$

$$=\int_{y=0\ldots H}\left[y^3/(3H^2)+(H-y)^3/(3H^2)\right]dy$$

$$=H^2/12+(1/(3H^2))\int_{z=H\ldots0}z^3\,(-dz)\,dy=H^2/6.$$

$y > H$:

$$\int_{y=H\ldots1}\left[\int_{x=0\ldots H}(x/H)^2+\int_{x=H\ldots y}dx+\int_{x=y\ldots1}0\,dx\right]dy$$

$$=\int_{y=H\ldots1}[H/3+y-H]\,dy$$

$$=H(1-H)/3+\int_{0\ldots1-H}z\,dz$$

$$=H(1-H)/3+(1-H)^2/2$$

Therefore:

$$E(CRPS(F,y))=H^2/6+H(1-H)/3+(1-H)^2/2$$

If $X \sim Unif[L, H]$, $0 \leq L \leq H \leq 1$, then the same method of calculation applies mutatis mutandis. If $L = 1 - H$ then the contributions from $x \leq y$ and $y \leq x$ are equal and we need only double the contribution from $x \leq y$. In that case, for $y \leq L$, the contribution from $x \leq y$ is zero, since $x > L$. Therefore, we compute:

$$\int_{y=L\ldots H}\int_{x=L\ldots y}F(x)^2\,dx\,dy+\int_{y=H\ldots1}\int_{x=L\ldots y}F(x)^2\,dx\,dy$$

$$=\int_{y=L\ldots H}\int_{x=L\ldots y}(x-L)^2/(H-L)^2\,dx\,dy+\int_{y=H\ldots1}\left[\int_{x=L\ldots H}(x-L)^2/(H-L)^2\,dx+\int_{x=H\ldots y}dx\right]dy$$

$$=\int_{y=L\ldots H}(y^3/3)/(H-L)^2\,dx+\int_{y=H\ldots1}[(H-L)/3+(y-H)]\,dy$$

$$=(H-L)^2/12+(H-L)(1-H)/3+(1-H)^2/2$$

Adding the identical contribution from $y \leq x$ gives:

$$E(CRPS(F,y))=(H-L)^2/6+2(H-L)(1-H)/3+(1-H)^2$$

Some values are in Table SI-2

Table SI-2

H	E(CRPS)
1	0.166666667
0.9	0.17
0.8	0.18
0.7	0.196666667
0.6	0.22
0.5	0.25
0.4	0.286666667
0.3	0.33
0.2	0.38
0.1	0.436666667
0	0.5

L	H	E(CRPS)
0.05	0.95	0.1675
0.1	0.9	0.17
0.2	0.8	0.18
0.3	0.7	0.19666667
0.4	0.6	0.22
0.5	0.5	0.25

Note that the expected *CRPS* for $X \sim Unif[0, H]$, $H \geq 0.5$ is the same as that for $X \sim Unif[1 - H, H]$. Thus, $E(CRPS)$ for $X \sim Unif[0, 0.7] = 0.19\underline{6} = E(CRPS)$ for $X' \sim Unif[0.3, 0.7]$. An expert who believes X finds that *30%* of the realizations Y are impossible has the same expected *CRPS* as an expert who believes X' finds 60% of the realizations are impossible. This illustrates how the *CRPS* compensates loss of statistical accuracy with a gain in 'sharpness'.

1.4 Scoring rules for average probabilities

To avoid problems with scores for individual items, (Cooke 1991) introduced scoring rules for average probabilities. Let random variables $X_1, ..., X_n$ take outcomes in a finite set, O, let M_O be the set of probability measures on O, and let M_n be the set of probability measures on $X_1, ..., X_n$. For $\Pi \in M_n$, let π be the vector of average probabilities, that is, $\pi_i = (1 / n) \Sigma_{j=1, ..., n} \Pi(X_j = i)$. Let s be the observed relative frequency of outcomes for realization, $(X_1, ..., X_n) = (x_1, ..., x_n)$. A scoring rule for average probabilities assigns a number, R, to the pair (π, s). R is strictly proper (positively sensed) if:

$$for\ all\ \Pi \in M_n,\ argmax_{\varphi \in Mo} E_{\Pi}(R(\varphi, s) = \pi.$$

This says, whatever the expert's belief, Π, about $(X_1, ..., X_n)$, (s)he achieves the maximal expected score by stating the probability, π, over outcomes which corresponds to their average probabilities. The proofs are a bit more complicated because, "for all Π," goes over a much larger set than the argmax over M_o.

There is a representation theorem in Cooke (1991) for such rules. However, more useful in practice are rules which are asymptotically strictly proper as $n \to \infty$. These rules allow the product form in the CM (see below). The proofs invoke the assumption that the expert's belief Π is a product measure. Recent expositions of the definitions and proofs are in the SI of (Colson and Cooke 2017) and the SI of (Cooke 2018)

Familiarity with foundations teaches that the problem of combining experts' judgments is not a mathematical problem. The laws of probability even supplemented with Savage's axioms and the theory of proper scoring rules, will never tell us how to combine experts. The problem is more akin to finding an optimal design in engineering. For example, a bicycle obeys Newton's laws but doesn't follow from them. Any working design will involve features motivated by practicalities in addition to laws.

So let it be with measures of "spread". Traditional measures like the standard deviation and prediction intervals are unsuitable because they inherit the physical dimension of the underlying variables: changing from meters to kilometres changes the numbers. To compare spreads across variables with different physical dimensions we need a measure which is scale invariant. We also need it to be "tail insensitive" because the tails in expert judgment are poorly constrained. Finally, it must be "slow" in order to prioritize statistical accuracy. The theory of asymptotic proper scoring rules for average probabilities gives a product form of: *measure of statistical accuracy* × *measure of lack of spread*. Weights will be formed by normalizing such products. We want statistical accuracy, a very fast function, to dominate; therefore the measure of "lack of spread" must be slow. Relative information is familiar and fits the bill. It requires specifying "relative to what". This is as it should be because information in a distribution is always relative to another (background) distribution.[ii]

Since the background distribution must integrate to one, it must be "concentrated" somewhere. The uniform or log-uniform distributions are chosen as default backgrounds because they have no location parameter other than the endpoints of their support. However, we do need to specify the compact support of the background for each item. We do this by choosing the smallest interval containing all assessments and the realization, if available, plus a *k%* overshoot. If k is very large then all experts appear very informative relative to the background, and this tends to discount the role of information in determining weights. The CM dictates that the choice of uniform/log-uniform and k are choices of the analyst, and any such choices must be controlled by user parameters in the code. Thus, we can see that k must be very large to make much difference. Uniform vs log-uniform can have an influence; the guidance is that, if we reason about orders of magnitude, use log-uniform.

[ii] You can see this in the continuous from of the entropy integral. For a discrete distribution $P = p1, ..., pn$, the entropy is defined as $H(P) = -\Sigma pi\ ln(pi)$. To pass to the continuous version replace $\Sigma \to \int$ and $pi \to f(x)dx$. That gives $H(f) = -\int fx(dx)\ ln(f(x)dx)$, which is meaningless. You must use relative information $I(f \mid g) = \int f(x)dx\ ln(f(x)dx/g(x)dx)$ so that the dx's inside the ln cancel and $I(f \mid g) = \int f(x)dx\ ln(f(x)/g(x))$.

2. Combining Experts

The combination schemes considered here are all examples of a linear pool. That is, a "Decision Maker *(DM)* is formed by forming a convex combination of the experts' densities. Each expert *e* is assigned a non negative weight w_e , the weights sum to unity, and *DM's* density for item *i* , $f_{DM}(i)$ is $\Sigma_e w_e f_e(i)$, where $f_e(i)$ is the density of expert *e* for item *i.*

2.1 Weights based on scoring rules for average probabilities

The 'combined score, Cs, of the expert, *e*, is the dimensionless quantity*:*

$$Cs_e = C_e * Inf_e \tag{2.1}$$

A scoring rule is (asymptotically) strictly proper if an expert achieves their (long run) maximal expected score by, and only by, giving assessments corresponding to their true beliefs. That is, an expert maximizes their long run expected score by, and only by, ensuring that the probabilities, *p = (0.05, 0.20, 0.25, 0.25 0.20, 0.05),* correspond to their true beliefs. The theory of proper scoring rules tells us that (2.1) becomes an asymptotic proper scoring rule if it is augmented with a cut-off, α, on the calibration score such that an expert is unweighted if $C_e < \alpha$. The value α is like the significance level in simple hypothesis testing, but its purpose is different. The goal is to measure "goodness" with a strictly proper scoring rule.

A combination of expert assessments is called a 'decision-maker' (*DM).* All *DMs* discussed here are examples of linear pooling. The Classical Model is essentially a method for deriving weights in a linear pool. A "good probabilistic expert" corresponds to an expert with good calibration (i.e., high statistical likelihood; high p-value) and high information. We want weights which reward good experts, and which pass these virtues on to the DM.

The reward aspect of weights is very important. We could simply solve the following optimization problem: find a set of weights such that the linear pool under these weights maximizes the combined score of the DM. When solving this problem with real data, one finds that the weights do not generally reflect the performance of the individual experts. As we do not want an expert's influence on the *DM* to appear haphazard, and we do not want to encourage experts to game the system by tilting their assessments to achieve a desired outcome; we must impose a strict scoring rule constraint on the weighting scheme.

The scoring rule constraint requires that the combined score is multiplied by the indicator function, $1_\alpha(C_e \geq \alpha)$, which takes the value *1* if $C_e \geq \alpha$ and *0* otherwise*:*

$$w_{\alpha,e} = C_e * Inf_e * 1_\alpha(C_e \geq \alpha) \tag{2.2}$$

This says that the expert, *e*, is weighted only if their statistical accuracy is at least α. The resulting *DM* is a function of α*:*

$$DM_{\alpha,i} = \sum_{e=1,\ldots,E} w_{\alpha,e} f_{e,i} \ / \sum_{e=1,\ldots,E} w_{\alpha,e} \tag{2.3}$$

Scoring rule theory does not say what the value of α should be. In practice there are three ways for choosing α. The 'optimized' *DM* chooses α such that the resulting combined score of the *DM* is maximized. The optimized *DM* is DM_{α^*} where α^* maximizes $C(DM_\alpha) \times Inf(DM_\alpha)$. This typically leads to choosing an α so high that only one or two experts are weighted. The 'statistical threshold' *DM* chooses an α to distribute weight over experts with "acceptable" calibration (typically, $\alpha = 0.05$ or 0.01). The "inclusive" *DM* chooses an α so low that all experts are weighted. This is also termed the "*non optimized pw*" If this choice is made a posteriori, then this is not a strictly proper scoring rule. It is to be noted that unweighted experts still have influence over the intrinsic range, and their rationales are recorded.

These weights are termed global because the information score is based on all the assessed calibration items. A variation on this scheme allows a different set of weights to be used for each item. This is accomplished by using information scores for each item rather than the average information score:

$$w_{\alpha,e,i} = 1_\alpha(C_e \geq \alpha) * C_e * f_{e,i}\ ln\ (f_{e,i} \mid g_i) \tag{2.4}$$

For each α we define the 'item weight' DM_α for the item, *i*, as:

$$IDM_{\alpha,i} = \sum_{e=1,\ldots,E} w_{\alpha,e,i}\, f_{e,i}\ \ / \sum_{e=1,\ldots,E} w_{\alpha,e,i} \tag{2.5}$$

The same variation applies to the threshold *DM* and the inclusive DM.

Item weights are potentially more attractive as they allow experts to up- or down-weight themselves for individual items according to how much they feel they know about that item. "Knowing less" means choosing quantiles further apart and thus lowering the information score for that item. Of course, the good performance of item weights requires that experts can perform this up/down-weighting successfully. Anecdotal evidence suggests that item weights improve over global weights as the experts receive more training in probabilistic assessment. For both global and item weights, calibration dominates over information; information serves to modulate between equally well calibrated experts. Definitions and proofs of these scoring rule properties are found in Cooke (1991, 2018, Colson and Cooke 2017).

Optimizing the weights in (2.3) and (2.6) often causes experts to be unweighted, even those with good scores. Such experts are not "rejected"; unweighting simply means that their input is already captured by a smaller subset of experts. Their value to the whole study is brought out in studying the robustness of the optimal *DM* under the loss of experts and in determining the intrinsic range. Their rationales are always included in the study results.

2.2 Other weighting schemes; predictive performance and inverse variance

The weights discussed above are all 'performance based' (i.e., an expert's weight depends on their performance). Another performance-based weight is the 'point-predictive-performance' weight, or simply 'predictive-performance' weight. Each item, *i*, with observed value, O_i, is divided by the prediction, $P_{e,i}$, of the expert, e, to form the ratio, $R_{e,i} = O_i / P_{e,i}$, assuming $P_{e,i} > 0$. For each expert, the exponentiated mean and standard deviation of the logged values of $R_{e,i}$ running over the values of *i*, termed the geometric mean and geometric standard deviation, are the performance measures for expert. The predictive-performance *DM* is formed by taking a weighted combination of models' densities where the weights are proportional to the variance over *i* of $R_{e,i}$.

It is becoming popular, especially in the climate modelling community, to average the results of models. A mathematical justification for this operation is sometimes based on treating model predictions as unbiased estimators with an imputed error term (Rougier et al 2013). There are many variations of this approach, but the simplest is based on the theory of weighted least squares. Suppose that uncertain quantity, X_i is estimated by measurement Z_i, with error, e_i. Suppose e_i is normally distributed with mean zero and standard deviation σ_i. Then after observing $Z_i = z_i$, the Renyi conditional distribution of X_i is normal with mean z_i and standard deviation σ_i. For *N* models with independent unbiased normal errors, the distribution of X_i conditional on observing $Z_1 = z_1, \ldots, Z_N = z_N$ is normal with a mean and variance given by:

$$E\ (X_i \mid Z_1 = z_1, \ldots, Z_N = z_N) = \Sigma_{i=1,\ldots,N}\ z_i\, w_i \tag{2.6}$$

$$Var\ ((X_i \mid Z_1 = z_1, \ldots, Z_N = z_N) = 1 / (1 / \sigma^2_1 + \ldots, 1 / \sigma^2_N) \tag{2.7}$$

$$w_i = (1 / \sigma^2_N) / (1 / \sigma^2_1 + \ldots, 1 / \sigma^2_N) \tag{2.8}$$

The weights are proportional to the inverse-variance of each error term. Because of Renyi conditionalization the posterior mean and variance apply to the uncertain quantity, X_i, and not, as in standard statistical

treatments, to the maximum likelihood estimator of X_i (see Cooke and Wielicki, 2018, for a full discussion). The rest of the derivations are entirely standard, and this *DM* is termed the 'weighted least squares' DM. To apply this theory, we must impute a variance to the assessments of the model for each item. We may expect good performance from this model if experts with small variance also have high statistical accuracy, which is not generally the case in the expert data reviewed here.

3. Validation

Validating experts' uncertainty quantification and that of combinations of experts' distributions using calibration variables from the experts' fields receives continuing attention. This section extends the data from 49 cases in (Cooke et al 2021) with 9 new cases involving in total 615 experts and 693 calibration variables. Publications describing these applications are found in (Cooke et al 2021).

3.1 In-Sample Validation

Evaluating performance based combinations of experts' distributions on the same data used to initialize the performance based weighting is termed in-sample validation. If performance based combinations were not superior to performance blind combinations in-sample there would be little point in pursuing out-of-sample validation. *pw* denotes item specific performance weights, *ew* denotes equal weights. Of the *58* studies in Table SI-3, statistical accuracy (SA) of *pw* exceeds that of *ew* on *45* studies, information (inf) of *pw* exceeds that of *ew* on *56* studies, and the combined score of *pw* exceeds that of *ew* on *53* studies.

Table SI-3 In-sample validation, 58 studies 2006-2021.

nr experts	nr calib vbls	study	ew SA	ew inf	ew comb	pw SA	pw inf	pw comb
4	10	Arkansas	0.386	0.198	0.076	0.499	0.523	0.261
9	10	Arsenic D-R	0.061	1.095	0.067	0.036	2.739	0.098
5	10	ATCEP	0.124	0.247	0.031	0.244	0.376	0.092
7	11	BFIQ	0.425	0.294	0.125	0.692	0.573	0.397
12	12	Biol agents	0.413	0.244	0.101	0.678	0.661	0.448
10	10	Brexit food	0.114	0.274	0.031	0.550	0.838	0.461
12	10	burkina faso	0.290	0.444	0.129	0.394	1.135	0.448
20	10	CDC ROI	0.233	1.230	0.286	0.720	2.305	1.660
48	14	CDC_All	0.250	1.082	0.270	0.968	2.541	2.460
10	11	CO2em	0.638	0.238	0.152	0.615	0.361	0.222
5	10	CoveringKids	0.628	0.274	0.172	0.720	0.506	0.365
7	10	CREATE	0.061	0.207	0.013	0.314	0.298	0.094
14	10	CWD	0.474	0.930	0.441	0.683	1.325	0.905
4	7	Daniela	0.533	0.168	0.089	0.554	0.634	0.351
8	10	dcpn_fistula	0.059	0.622	0.037	0.266	1.343	0.357
14	15	eBBP	0.358	0.316	0.113	0.833	1.406	1.172
14	8	Eff_Erupt	0.286	0.796	0.228	0.664	1.240	0.823
11	15	Erie Carps	0.313	0.294	0.092	0.761	0.856	0.651
11	10	ethiopia	0.474	0.659	0.312	0.707	1.740	1.230
5	8	FCEP	0.222	0.099	0.022	0.664	0.574	0.381
7	10	Florida	0.756	0.455	0.344	0.756	1.145	0.866
5	10	France	0.078	0.433	0.034	0.652	1.958	1.276
9	11	GDP2300	0.370	0.266	0.098	0.706	0.673	0.475
8	18	GeoPol	0.196	0.559	0.109	0.425	1.150	0.488
12	14	Gerstenberger	0.644	0.482	0.310	0.756	1.202	0.909
9	13	GL-NIS	0.044	0.307	0.014	0.928	0.259	0.240

6	10	Goodheart	0.550	0.277	0.153	0.707	0.959	0.678
18	8	Hemophilia	0.254	0.202	0.051	0.312	0.463	0.144
20	16	ICE_2018	0.128	0.545	0.070	0.942	0.928	0.875
10	11	Ice_2012	0.492	0.517	0.254	0.615	1.038	0.639
5	10	Illinois	0.620	0.264	0.163	0.386	0.599	0.231
8	11	IQEarn	0.705	0.575	0.405	0.705	0.623	0.439
4	10	Italy	0.218	0.197	0.043	0.447	0.466	0.209
11	14	Leontaris	0.039	0.132	0.005	0.968	0.380	0.368
11	10	liander	0.228	0.484	0.111	0.683	0.751	0.513
4	10	Nebraska	0.368	0.695	0.256	0.033	1.447	0.048
5	10	Nogal	0.114	0.278	0.032	0.290	0.496	0.144
4	10	obesity	0.070	0.243	0.017	0.780	0.490	0.382
6	30	Peyras30	0.103	0.118	0.012	0.063	0.597	0.038
10	13	PHAC_T4	0.265	0.204	0.054	0.096	0.492	0.047
16	21	PoliticalViolence	0.443	1.047	0.463	0.129	1.818	0.234
9	13	Puig-GDP	0.063	0.435	0.027	0.928	0.992	0.920
8	20	Puig-oil	0.881	0.201	0.177	0.128	0.614	0.079
9	10	rwanda	0.474	0.322	0.152	0.683	0.861	0.588
8	10	San Diego	0.334	1.066	0.356	0.345	1.186	0.409
14	15	Sheep	0.661	0.780	0.516	0.643	1.310	0.843
5	10	Spain	1.22E-05	0.231	2.82E-06	3.59E-05	0.690	2.47E-05
14	16	SPEED	0.517	0.751	0.389	0.992	0.783	0.777
12	13	Tadini Clermont	0.329	0.280	0.092	0.755	1.144	0.863
8	13	Tadini Quito	0.421	0.232	0.098	0.928	0.849	0.788
18	17	TdC	0.166	0.364	0.060	0.989	1.256	1.242
7	10	tobacco	0.200	0.451	0.090	0.688	1.062	0.730
21	16	Topaz	0.629	0.922	0.580	0.411	1.455	0.598
6	10	UK	0.132	0.331	0.044	0.218	0.661	0.144
9	11	umd	0.068	0.804	0.054	0.706	1.988	1.404
32	18	USGS	0.058	0.795	0.046	0.507	1.512	0.766
5	10	Washington	0.155	0.529	0.082	0.499	0.988	0.493
12	8	rijn_faalkansen	0.534	0.669	0.357	0.688	2.100	1.444

3.2 Out-of-Sample validation

Out-of-sample validation occurs when performance is evaluated on a set of variables which is disjunct from the variables used to initialize the models. Since the variables of interest are seldom observed in the time frame of the studies, out-of-sample validation reduces to cross-validation: The calibration variables are split into training and testing sets. The combination models are initialized on the training set and performance is evaluated on the test set. Difficulties involved in cross validation discussed in (Colson and Cooke 2016) include: (i) inability to include optimization in performance based combinations, (ii) selecting suitable training/testing splits, (iii) aggregating results over diverse studies and (iv) excessive computing times. The hypothesis that *pw* and *ew* combinations were statistically indistinguishable was rejected at the *1.8E-7* level in the most recent results (Cooke et al 2021).

3.3 Random Expert Hypothesis

Issues with cross-validation prompted a new approach to validating expert uncertainty quantification (Cooke et al 2021). The Random Expert Hypothesis (REH) states that putative differences in performance between

experts are just noise and do not indicate persistent differences among the experts. One way to test this hypothesis is to compare panel wide performance metrics in the original panel with the same metrics as generated by a large set of "scrambled panels" in which the assessments are randomly re-allocated to experts, thus wiping out any 'expert effect'. The code used for this exercise is based on the *5%, 50%* and *95%* quantiles. In a preliminary study (Marti et al 2021), including the *25%* and *75%* values had no effect on REH. Note that REH is implicitly assumed in all performance-blind combination schemes. Note also that testing REH avoids the intermediary of constructing combinations of experts' judgments and avoids splitting the calibration variables. Note finally that if all experts are "equally good" or "equally bad" in terms of statistical accuracy and information, then REH would actually be true. REH fails if the differences in expert performance are greater than that which random scrambling can re-produce.

We are interested in the panel averages, standard deviations, maxima and minima over the 58 studies of the combined scores and also for the most important component of the combined score, statistical accuracy. The distribution of these metrics based on 1000 scrambled panels represents the variation in these metrics which would result if differences in expert performance were due to noise. If REH were true then the metrics in the original panel could just as well be drawn randomly from the scrambled distributions. We thus consider the percentiles in the scrambled distributions which are realized by the metrics in the original panel. If REH were true, the realized percentiles should be uniformly distributed on *[0,1]*. For most metrics, high inter-expert differences correspond to high percentiles in the scrambled distributions. For minima, we plot *1 – percentile* so that high plotted values indicate differences in the original panel which scrambling has difficulty reproducing and which are favourable to performance weighting. Figure SI-1 plots the results; it is evident that the original metrics are not uniformly distributed in the scrambled panels.

The hypothesis *the percentiles of a metric of the original panel realized in the scrambled distribution are independent across the 58 studies, and the probability for a value above 0.5 is ½,* is the subject of the *binomial test*. This test ignores the size of the departures from *0.5*. The levels at which REH is rejected for the various metrics are comparable to that reported in the previous section for cross-validation.

The sum of the *58* percentiles for *Av SA* is *44*. Since the sum of *58* independent uniform variables is very nearly normal with mean *58/2* and variance *58/√12*. The probability of exceeding *44* is *1.5 E-12*. This is termed the *sum test* and it is much more powerful than the binomial test for all metrics, as shown in Table SI-3**.**

The percentile for *Av SA* in the study BFIQ is *0.795*. If we reject REH for this study, we have probability *1 – 0.795* that REH actually holds for BFIQ. If we sum the probabilities for false rejection over all studies we find the expected number of false rejections in the 58 studies. Table SI-4 shows the percentage of false rejections for each of the metrics.

An interesting feature of Table SI-4 is that the panel minima for SA and for the combined scores yield the lowest P-values and the smallest number of expected false rejections. This means that the random scrambling has more difficulty generating scores lower than these minima in the original panels, than for the other metrics. Overall, if we reject REH for each study we may expect that between one fourth and one fifth of the studies REH may be true, depending on the chosen panel metric.

Figure SI-1 Percentiles of panel metrics realized by the original panel in the distribution of scrambled panels

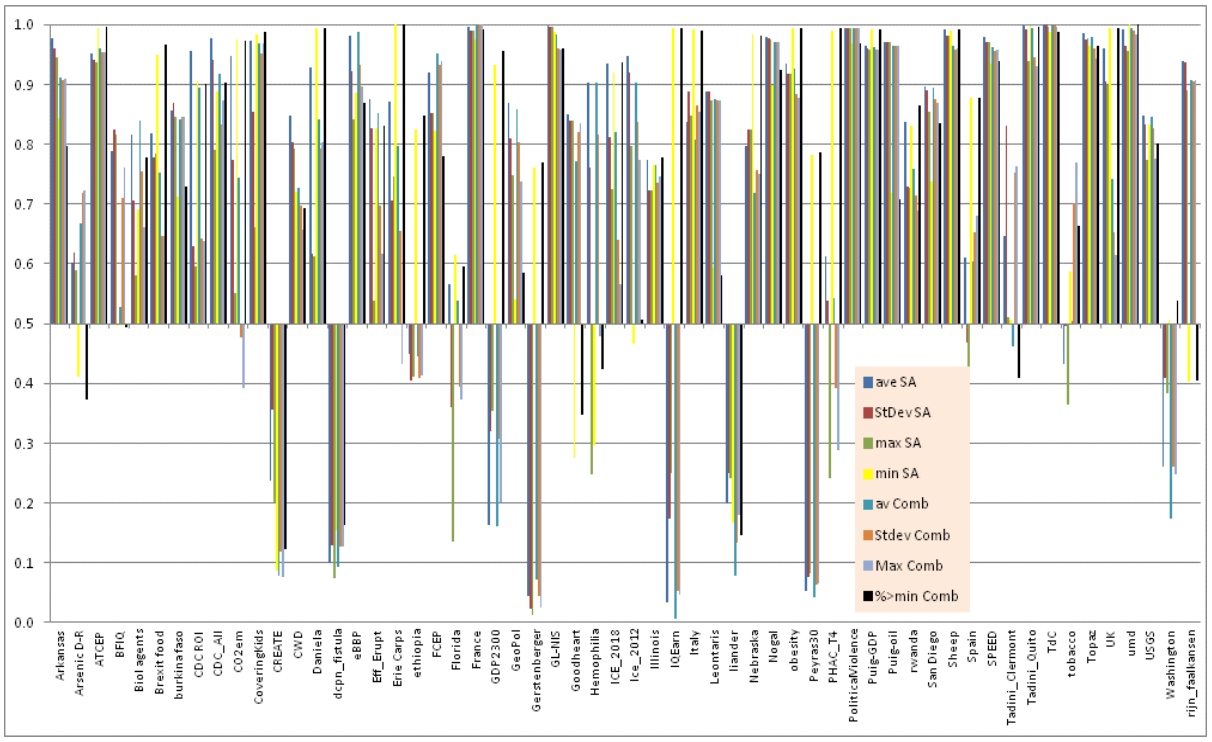


Table SI-4 Test results for REH for the binomial and sum tests.

	ave SA	StDev SA	max SA	%>min SA	av Comb	Stdev Comb	Max Comb	%>min Comb
#>0.5	48	46	44	49	48	46	44	49
Binom test	2.26E-07	4.11E-06	5.02E-05	4.48E-08	2.26E-07	4.11E-06	5.02E-05	4.48E-08
Sum	44.57	42.92	39.55	44.93	42.25	41.23	39.95	45.37
Sum test	7.14E-13	1.23E-10	7.93E-07	2.18E-13	8.48E-10	1.34E-08	3.18E-07	4.76E-14
Exp'd % false rejections	23.16%	26.01%	31.81%	22.54%	27.16%	28.92%	31.12%	21.77%

Supplemental References

Colson AR, Cooke RM. 2017 Cross validation for the classical model of structured expert judgment. Reliability Engineering & System Safety 163, 109–120. (doi:10.1016/j.ress.2017.02.003)

Cooke R.M., Marti D, Mazzuchi T. 2021 Expert forecasting with and without uncertainty quantification and weighting: What do the data say? International Journal of Forecasting 37, 378–387. (doi:10.1016/j.ijforecast.2020.06.007)

Cooke RM, Wielicki B. 2018 Probabilistic reasoning about measurements of equilibrium climate sensitivity: combining disparate lines of evidence. Climatic Change 151, 541–554. (doi:10.1007/s10584-018-2315-y)

Cooke RM. 1991 Experts in Uncertainty: Opinion and Subjective Probability in Science. Oxford, UK: Oxford University Press.

Cooke, Roger M. (2014) Validating Expert Judgments with the Classical Model" in Experts and Consensus in Social Science - Critical Perspectives from Economics, Sociology, Politics, and Philosophy. Editors: Carlo Martini and Marcel Boumans, Series title: Ethical Economy - Studies in Economic Ethics and Philosophy, Springer.

Cooke R.M., (2018) Validation in the Classical Model, (pp 37-59), Online Appendix SI Strictly Proper Scoring Rules as Weights Dias LC, Morton A, Quigley J (eds), Elicitation: The science and art of structuring judgement. Springer, New York, 2018.

de Finetti B. 1937 La prévision : ses lois logiques, ses sources subjectives. Annales de l'institut Henri Poincaré 7, 1–68.

DeGroot MH, Fienberg SE. 1983 The comparison and evaluation of forecasters. The Statistician 32, 14–22.

Marti, H.D., Mazzuchi, T.A. and Cooke R/.M. (2021) Are Performance Weights Beneficial? Investigating the Random Expert Hypothesis, in Expert Judgement in Risk and Decision Analysis eds Nane, Hanea, French and Bedford, Springer Nature Switzerland AG, Cham, Switzerland

Murphy AH. 1977 The value of climatological, categorical, and probabilistic forecasts in the cost-loss ratio situation. Monthly Weather Review 105, 803–816.

Ray, Evan L., Nutcha Wattanachit, Jarad Niemi, Abdul Hannan Kanji, Katie House, Estee Y Cramer, Johannes Bracher, Andrew Zheng, Teresa K Yamana, Xinyue Xiong, Spencer Woody, Yuanjia Wang, Lily Wang, Robert L Walraven, Vishal Tomar, Katharine Sherratt, Daniel Sheldon, Robert C Reiner Jr, B. Aditya Prakash, Dave Osthus, Michael Lingzhi Li, Elizabeth C Lee, Ugur Koyluoglu, Pinar Keskinocak, Youyang Gu, Quanquan Gu, Glover E. George, Guido España, Sabrina Corsetti, Jagpreet Chhatwal, Sean Cavany, Hannah Biegel, Michal Ben-Nun, Jo Walker, Rachel Slayton, Velma Lopez, Matthew Biggerstaff, Michael A Johansson, Nicholas G Reich, (2020), Ensemble Forecasts of Coronavirus Disease (COVID-19) in the U.S. medRxiv 2020.08.19.20177493; Posted August 22, 2020 doi: https://doi.org/10.1101/2020.08.19.20177493

Rougier J, Goldstein M, House L. 2013 Second-Order Exchangeability Analysis for Multimodel Ensembles. Journal of the American Statistical Association 108, 852–863. (doi:10.1080/01621459.2013.802963)

Shuford EH, Albert A, Edward Massengill H. 1966 Admissible probability measurement procedures. Psychometrika 31, 125–145. (doi:10.1007/BF02289503)