Expert Judgment and the Inevitability of Validation

Roger M. Cooke, Deniz Marti, Thomas Mazzuchi

Sept 2, 2019

Abstract

We present latest results in validating Structured Expert Judgment. The post-2006 expert judgment data base is extended to 49 studies. In-sample and out-of-sample results are updated with the latest studies. The assumption underlying all performance-blind combination schemes is the *Random Expert Hypothesis*: putative differences in expert performance are due to random stressors and are not persistent properties of the experts themselves. Using methods of randomly scrambling expert panels developed in previous work, we generate distributions for a full set of performance metrics. The hypotheses that the original panels' performance values are drawn from distributions produced by random scrambling are rejected at the significance level of E-12. Random stressors cannot produce the variation in performance seen in the original panels. Rejecting the random expert hypothesis for all 49 studies would lead to 10 expected false rejections. Aggregating over all variables and all studies we compare prediction errors relative to the prediction error of an equal weighted combination of experts' medians. Using the medians of performance weighted combinations yields a 58% improvement. Using a performance weighted combination of experts' medians yields a 43% improvement. Using the medians of equally weighted combinations of experts distributions yields a 29% improvement. The prediction error caused by combining quantiles instead of combining distributions (as has been incautiously proposed), is greater for equal weighting than for performance weighting.

Introduction:

Using expert uncertainty quantification (UQ) as scientific data with traceability and validation dates from (Cooke et al., 1988; Cooke, 1987; Cooke, 1991) under the name "Classical Model" or "Structured Expert Judgment". Distinguishing features are treating experts as statistical hypotheses and evaluating performance with respect to statistical accuracy and informativeness based on calibration variables (a.k.a. seed variables) from their field to which true values are / become known. Combinations of experts' distributions (termed decision makers or DMs) using performance-based (PW) and equal weighting (EW) are compared based on performance on calibration variables.

The expert data up to 2006 was made publically available in 2008 (Cooke & Goossens, 2008) and is currently available at http://rogermcooke.net/. The best current summary of the Classical Model, applications and validation research are published by Colson and Cooke (2017 and 2018; also see online supplements).The reader is referred to these sources for older publications.

Application highlights involved nuclear safety in the 1990s with the European Union and the United States Nuclear Regulatory Commission, fine particulates with Harvard University and the government of Kuwait in 2004-2005, food-borne diseases for the World Health Organization (WHO) in 2011-2013, ice sheet dynamics for Princeton University and Rutgers University and Resources for the Future (Bamber et al., 2019), and volcanic hazard levels in different parts of the world. The Classical Model was a key decision-support procedure during the prolonged eruption on the island of Montserrat, West Indies in 1995-2018. Over the same period, expert elicitations using the Classical Model have informed many issues of public health policy, civil aviation safety, fire impacts on engineered structures, and earthquake resistance of critical utility facilities.

Validation is the cornerstone of science. A special issue on expert judgment (Cooke & Goossens, 2008) focused on this issue. Colson and Cooke (2017) gave an extensive review of validation research and applied the cross validation code of Eggstaff et al. (2014) to the 33 professional studies conducted post-2006. These studies are more uniform in design and better resourced than the earlier studies. 320 experts produced in total 3,777 expert assessments of calibration variables. Those numbers have since grown to 49 studies involving 516 experts and 6,508 expert assessments of calibration variables.

The driver behind virtually all these applications is the validational aspect of the Classical Model. Although the psychological community has long drawn attention to cognitive biases inherent in expert UQ, and despite a robust interest in validation research, there are barriers to the use of performance measures. The conclusion speculates on possible explanations for this. Section 2 updates the data on expert performance and in-sample validation, Section 3 updates results on cross validation and Section 4 focuses on a new direction termed the Random Expert Hypothesis (REH). Section 5 explores co-benefits of performance weighting in terms of point predictions. Section 6 gathers conclusions.

2. Post-2006 Expert Data and In-sample Validation

In addition to the 33 post-2006 studies studied in Colson and Cooke (2017; 2018), 16 more studies have been completed as post-2016, as summarized in Table 1 (see Appendix A for data references).

Table 1 Expert judgment studies are illustrated with the number of calibration variables and experts, post-2006 (post-2016 bolded)

Study	# of Experts	# of Calibration Variables	Subject
UMD	9	11	Nitrogen removal in Chesapeake Bay
arsenic	9	10	Air quality levels for arsenic
Biol Agents	9	10	Human dose-response curves for bioterror agents
ATCEP	5	10	Air traffic Controllers Human Error
Daniela	4	10	Fire prevention and control
eBBP	14	15	XMRV blood/tissue infection transmission risks
create	7	10	Terrorism
effErupt	14	8	Icelandic fissure eruptions: source characterization
erie	10	15	Establishment of Asian Carp in Lake Erie
FCEP	5	8	Flight Crew Human Error
Sheep	14	15	Risk management policy for sheep scab control
hemophilia	18	8	Hemophilia

Liander	11	10	Underground cast iron gas-lines
PHAC	10	12	Additional CWD factors
TOPAZ	21	16	Tectonic hazards for radwaste siting in Japan
SPEED	14	16	Volcano hazards (Vesuvius & Campi Flegrei, Italy)
TDC	18	17	Volcano hazards (Tristan da Cunha)
GL	9	13	Costs of invasive species in Great Lakes
Goodheart	5	10	Airport safety
Ice	10	11	Sea level rise from Ice Sheets melting due to global warming
YTBID (CDC)	14	48	Return on investment for CDC warnings
Gerestenberger	12	13	Probabilistic Seismic-Hazard Model for Canterbury
CWD	14	10	Infection transmission risks: Chronic Wasting Disease from deer to humans
Nebraska	4	10	Grant effectiveness, child health insurance enrollment
SanDiego	7	10	Effectiveness of surgical procedures
Arkansas	4	10	Grant effectiveness, child health insurance enrollment
Covering Kids	5	10	Grant effectiveness, child health insurance enrollment
dcpn_Fistula	8	10	Effectiveness of obstetric fistula repair
Florida	7	10	Grant effectiveness, child health insurance enrollment
Illinois	5	10	Grant effectiveness, child health insurance enrollment
Obesity	4	10	Grant effectiveness, childhood obesity
Tobacco	7	10	Grant effectiveness, childhood obesity
Washington	5	10	Grant effectiveness, child health insurance enrollment
cdc-roi	**20**	**10**	**Return on investment for CDC warnings**
IQ-earn	**8**	**11**	**Effects of Increases in IQ in India on the Present Value of Lifetime Earnings**
Brexit_food	**10**	**10**	**Food price change after Brexit**
TadiniQuito	**8**	**13**	**Somma-Vesuvio volcanic complex geodatabase**
Tadini_Clermont	**12**	**13**	**Somma-Vesuvio volcanic complex geodatabase**
PoliticalViolence	**16**	**21**	**Political Violence**
ICE_2018	**20**	**16**	**Future see level rise**
Geopolit	**9**	**16**	**Geopolitics**
puig-gdp	**9**	**13**	**Emission forecasts from Mexico**

puig-oil	**6**	**19**	**Oil emissions and prices**
France	**5**	**10**	**Future Antimicrobial Resistance in France**
Italy	**4**	**8**	**Future Antimicrobial Resistance in Italy**
Spain	**5**	**10**	**Future Antimicrobial Resistance in Spain**
UK	**6**	**10**	**Future Antimicrobial Resistance in UK**
USGS	**18**	**32**	**Volcanos**
BFIQ	**7**	**11**	**Breastfeeding and IQ**

Note. Post-2016 studies are bolded in text.

The histogram of calibration variables and the graph of *p* value scores are given in Figure 1.

Figure 1: Calibration frequencies (left) for all 516 post-2006 experts and *p* value scores (right) not accounting for differences in statistical power. The traditional 5% threshold for simple hypothesis testing is given as a red line.

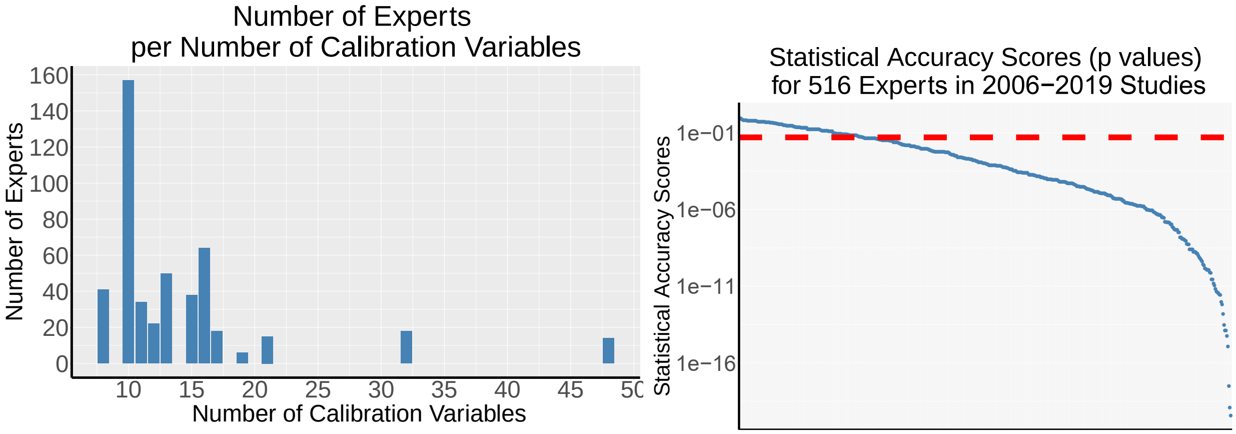
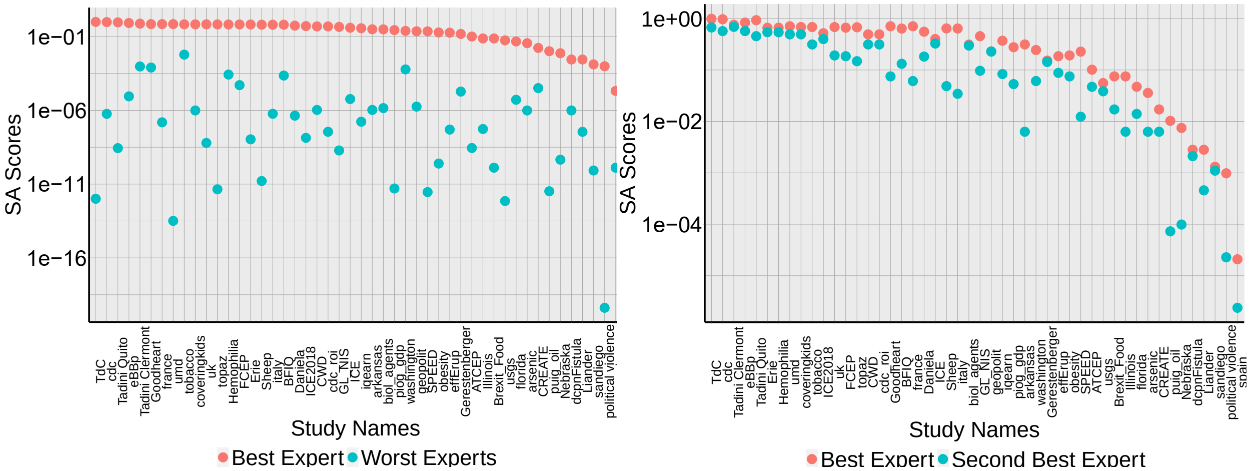


Figure 2 shows the best and worst *p* values (statistical accuracy scores) per study for all post-2006 studies and the two best performing experts. There are generally 4 or more orders of magnitude in statistical accuracy scores between the best and worst expert per study. Despite the fact that only 133 of the 516 experts would not be rejected as statistical hypothesis at the 5% level on simple hypothesis tests, most studies have one or even two statistically acceptable experts.

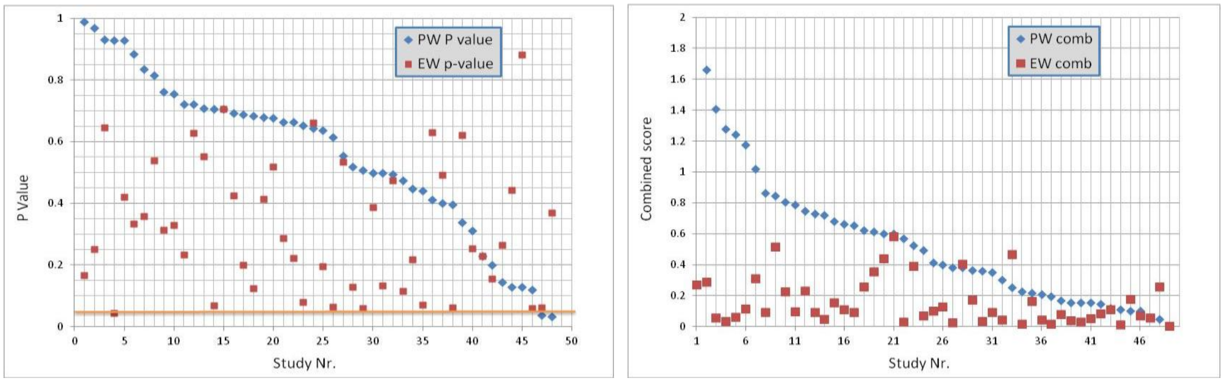
Figure 2: Best and worst expert (BEP and WEP, respectively) *p* values or Statistical Accuracy (left) and *p* values of best two experts in terms unnormalized weight (combined score of statistical accuracy × informativeness) (right), per study, 2006-2019 data.



Note. Studies are ordered with respect to best experts' statistical accuracy scores in both plots.

Comparing PW and EW decision makers on the data used to initialize the performance weighting is "in-sample validation". Figure 3 shows the in-sample results for the 2006-2019 data. The combined score is the product of the statistical accuracy and the informativeness score. The left graph shows that PW and EW have roughly the same number of studies below the conventional 5% rejection threshold, though PW tends to be higher. The right graph adds the informativeness factor and boosts the in-sample superiority of PW over EW.

Figure 3: In-sample comparison of performance weighted (PW) and equal weighted (EW) decision makers with respect to *p* values (left) and combined scores (right), ordered by PW values.

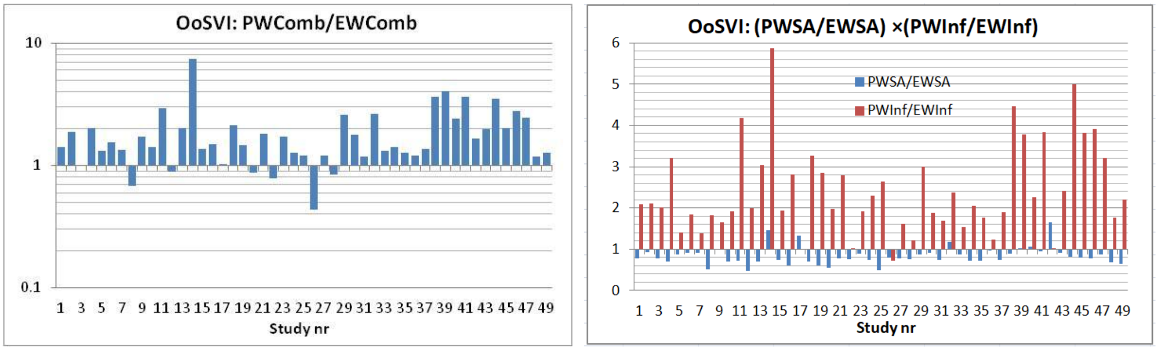


3. Out-of-sample; Cross Validation

Unless the variables of interest can be observed shortly after completion of a study, out-of- sample validation comes down to cross validation. The calibration variables are split into a training set for initializing the PW and a test set for comparing PW and EW . The sets on which the performance weights are derived and evaluated are thus disjoint.

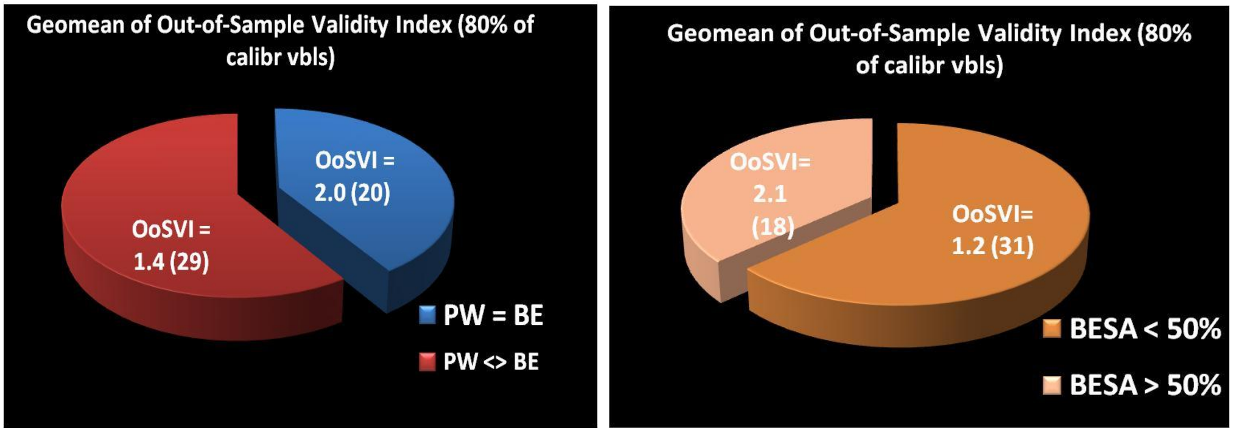
Many issues involved in choosing the training and test set sizes are discussed in Colson and Cooke (2017), to which we refer the interested reader. The upshot is that using 80% of the calibration variables as a training set best balances the competing goals of resolving expert performance on the training set and resolving the performance of combinations on the test set. The training set then has enough statistical power to reduce the variance in the expert weights thereby rendering the performance weights similar to the weights based on all calibration variables. The test set loses statistical power in resolving the PW and EW DMs, but with 10 calibration variables statistical accuracy scores for assessments of 5th , 50th ,and 95th percentiles still vary by a factor 31. Moreover, higher resolution is of no value if the PW DM is very volatile and unlike the PW DM of the full study. Of course the actual sizes of the training and test sets vary with the total number of calibration variables. The 80% split makes it easier to pool the results from all studies. With 10 calibration variables, there are 45 distinct 8-tuples of calibration variables to be used as training sets. Performance is scored on the 2 remaining variables. The statistical accuracy, the informativeness and the combined score (the product of the former two) are averaged over the 45 different test sets. In Colson and Cooke (2017), it is shown that the average ratio of combined scores for PW and EW is indistinguishable from, the ratio of average combined scores for fixed training set size. The ratio of combined scores based on 80% of the calibration variables is called the "Out of Sample Validity Index (OoSVI)."

Figure 4: Ratios of combined scores PW / EW averaged over all training sets sized at 80% of the calibration variables for 49 post-2006 studies (left). Combined scores factored into ratios of statistical accuracy (SA) and informativeness (Inf) (Right).



For 42 of the 49 studies the ratio PWcomb / EWcomb is greater than 1. Under the null hypothesis that there is no difference between PW and EW, the probability of seeing 42 or more ratios greater than 1 is 1.2E–8. The right panel of Figure 4 shows that PW suffers a modest out-of-sample penalty in statistical accuracy, which is more than compensated by a boost of informativeness. The mean statistical accuracy score (i.e., *p* value) for EW is 0.54, while that of PW is 0.43.

Figure 5: Breakdown of OoSVI scores in Figure 4 according to whether PW was identical with the best expert (PW=BE; left) and whether BE's statistical accuracy was greater than 50% (BESA>50%; right).



As in Colson and Cooke (2017), the features which best explained the differences in OoSVI were studied. The results echo those earlier findings; if PW concentrates all weight in the best expert (BE) the overall geomean of OoSVI for all studies (1.63) splits into 2.0 (PW=BE) and 1.4 (PW ≠ BE). Similar results are obtained by splitting into studies in which BE's statistical accuracy is above (2.1) resp. below (1.2) 50%. Other features such as number experts, number of calibration variables and plenary versus individual elicitation had less effect. The quality of the best expert is the main determinant for OoSVI. The rank correlation between OoSVI and the in sample ratio of combined scores is 0.5; these measures are related but not identical for reasons addressed in the following paragraph.

Cross validation is essential for demonstrating the value of PW relative to EW for out-of-sample prediction. The necessity of splitting the calibration set into training and test sets exacts a toll, which is illustrated with the "Ice Sheet 2018" study ('ICE_2018;' see Table 1) including 20 experts and 16 calibration variables. With a training set of 13 (80% of 16), there are 560 distinct training sets. 8 of the 20 experts were weighted on at least one of these sets. For 7 of these 8, the difference between their maximal and minimal weight was 1; that is their weights vacillated between 0 and 1. The PW combinations evaluated on the 3 test variables still exhibit volatility and deviate from the PW of the entire study. The most we can say is that the OoSVI compares the score of EW with the scores of a swarm of PW, which loosely resemble the PW of the full study. The cross validation data validates the performance weighting method, but not a specific PW combination.

4. The Random Expert Hypothesis

Recently a new approach to validation has emerged (Marti et al., 2019). Whereas cross validation is hampered by the two-sided loss of statistical power caused by splitting calibration variables into training and test sets, the new approach does not focus on the performance of a combination of experts. Instead, it focuses on the performance of the experts themselves and investigates the assumption underlying all performance-blind approaches; namely that performance measures are unable to make meaningful distinctions in experts' performances. This may be because the experts are all equally good or equally bad. It may also be that any putative differences are swamped by the noise inherent in expert judgment: experts are influenced by random stressors, they have good or bad days, their performance is affected by the particular choice of calibration variables, etc. The claim that putative differences in expert performance cannot be statistically distinguished from random fluctuations is called the Random Expert Hypothesis (REH). It is not defended with empirical arguments but is invoked (albeit implicitly) by any performance-blind approach. Without performance data, the REH can never be dislodged and the advocates of performance-blindness seemingly carry no proof burden other than raising doubts about claims of PW superiority (Winkler et al., 2018).

The cross validation work sketched above does constitute a rebuttal of the REH, but the rebuttal is based on the claim that PW outperforms EW out-of-sample. This does not test the REH directly. The low power of the test set means that the DM's statistical accuracy scores are poorly resolved. The new approach to validation focuses directly on REH without the intermediary of performance based combinations. Indeed, the REH itself carries a heavy proof burden and can be tested using expert performance provided by our 49 post-2006 studies. Intuitively, if performance differences are the result of noise, then randomly re-allocating the experts' assessments among the panel members will randomly re-distribute the random stressors. The fluctuations in performance produced in this way should envelope those in the original panel. For example, the maximum scores in the original panels should resemble the maxima in a large set of randomly generated panels. If not, then having the best score must be a persistent property of the expert in question and not the result of random fluctuations. It emerges that tests of REH are much more powerful than the cross validation tests.

To make this idea precise, consider a *random scramble* of an expert panel composed of 15 experts and 10 calibration variables. We create 'scrambled expert 1' by randomly choosing an assessment without replacement from one of the 15 experts for the first variable, a second random draw without replacement gives the second assessment for 'scrambled expert 1' and so on. 'Scrambled expert 2' chooses his assessments in a similar way from the assessments not chosen by 'scrambled expert 1'. The final scrambled expert, 'scrambled expert 15' gets the leftovers. In this scrambled panel, we can measure the statistical accuracy (SA) and informativeness (Inf) of each expert, we can measure the combined scores (SA × Inf), we can compute the average scores, the maximum and minimum and the standard deviation of the scores.

For each study, we repeat the scrambling 1000 times and build up a distribution for each performance metric. This distribution reflects the variation we should see in that performance metric if experts' performance differences really were due only to random stressors. Suppose we compute the average SA for experts in each of the 1000 scrambled panels for a given study. The REH now asserts that the average SA in the original panel looks like it is drawn from this distribution. There should be a 50% chance that the original SA is above the median of the REH distribution, a 5% chance that it is above the 95th percentile of the REH, etc. Thus, REH expects that in 2.45 of the 49 studies, the original average SA should fall above the 95th percentile of the REH distribution. Actually, this happens in 20 of the 49 studies. The probability of 20 or more studies falling above the 95th percentile if REH were true is 6.5E−14. REH fails if the differences in the experts themselves in the original panel are greater than what can be produced by scrambling the experts.

Note that random scrambling will have no effect on the EW combination. This underscores the fact that EW implies REH. In consequence (modus tollens), if REH is (statistically) rejected, then so is EW. In this sense, REH provides a more powerful test of the assumption underlying the use of EW. The same holds for the "averaging quantile" approaches (Lichtendahl et al., 2013) or indeed any approach which is performance-blind. If all experts in a panel are "equally good" or "equally bad," then REH may actually be true for that panel, which indeed could sometimes happen. The use of PW depends on the fact that such panels are in the minority. Testing REH on a set of cases allows us to gauge the size of that minority.

The data has been standardized in ways that do not affect the REH; experts who did not assess all calibration variables were dropped, all background measures are converted to uniform and only the 5th , 50th and 95th percentile elicitations were used (Marti et al., 2019).

For each of the 49 studies, the following seven performance metrics shown in Figure 6 are computed for the original panel and for each of the 1000 scrambled panels:

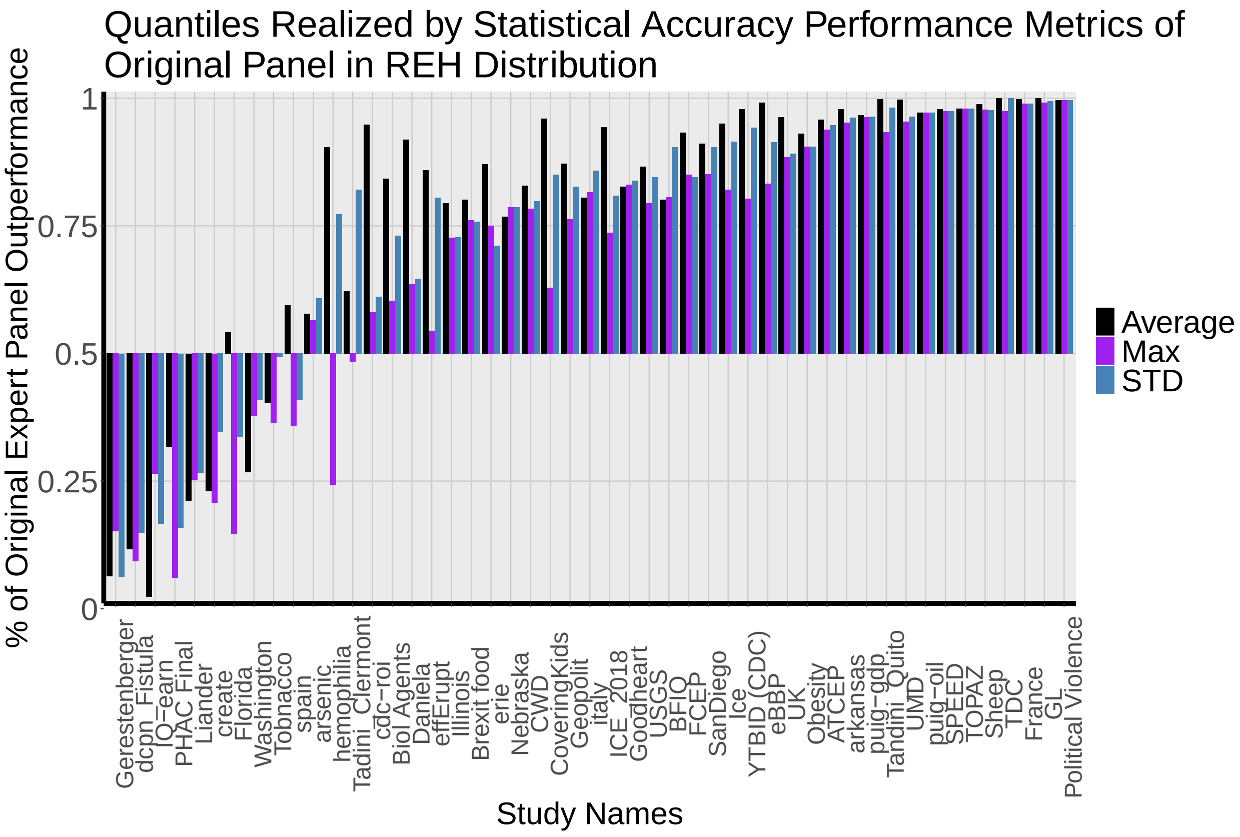
1. Panel Average Statistical Accuracy
2. Panel Max Statistical Accuracy
3. Panel Standard Deviation of Statistical Accuracy
4. Panel Average Combined Score
5. Panel Max Combined Score
6. Panel Standard Deviation of Combined Score
7. Panel Min Combined Score

For the first six metrics, we are interested in the quantile of the REH distribution realized by the metric in the original panel. For the last metric we are interested in the complimentary quantile, that is, we are interested in the percentage of the 1000 scrambled panels in which the original minimum is *lower* than the scrambled minimum. This is done so that all metrics have the same sense: higher is better for PW, and worse for EW.

If REH were true, that is, if the original panel's metrics were really drawn from the REH distribution, then the quantiles in Figure 6 should be uniformly distributed on the interval [0, 1]. The number of bars above the value 0.5 should be statistically equal to the number below 0.5. The "amount of color" above 0.5 should statistically equal the amount below 0.5.

There are two simple tests for the REH hypothesis. The binomial test simply counts the number of values greater than 0.5 for each metric and reports the *p* value for the corresponding null hypothesis: *the probability that 50% of the random panels outperform the original panel metric is 0.5*. The binomial test does not consider how far above or below 0.5 the metrics are. The *sum test* simply adds the 49 quantiles for each metric. Under REH this sum should be (very, very nearly) normally distributed with mean 49/2 = 24.5 and standard deviation $(49/12)^{½}$ = 2.02. For example, 'Average Statistical Accuracy' in the original panel exceeds the median of the REH distribution for 'Average Statistical Accuracy' on 42 of the 49 studies. If the probability of exceeding the median were really 0.5, the probability of seeing 42 or more "successes" would be 1.81 E-7. Summing the original panels' 49 realized quantiles in the REH distribution for 'Average Statistical Accuracy' yields 38.36. The probability that a normal variable with mean 24.5 and standard deviation 2.02 exceeds 38.36 is 3.48 E-12. The sum test is much more powerful than the binomial test. Table 2 collects the results for the binomial and sum tests. Suppose we reject REH for each of the 49 studies. The sum of the *p* values gives the expected number of false rejections. This number is 49 – 38.36 = 10.64. The expected percentage of studies in which REH would be falsely rejected is thus 22%.

Figure 6: Quantiles of the REH distributions realized by the performance metrics for Statistical Accuracy (top) and Combined Score (bottom) by the original expert panels, for 49 studies.



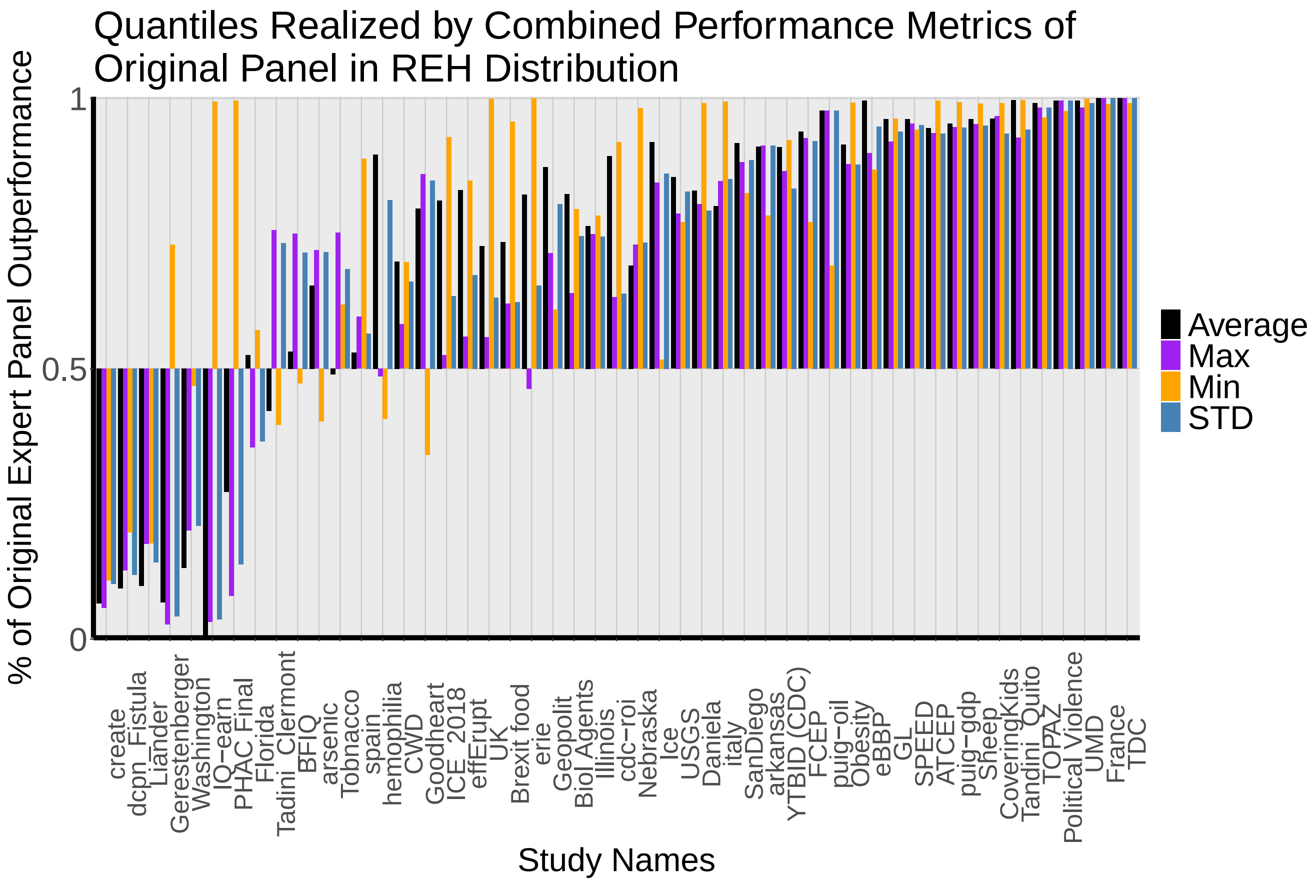


Table 1: *p* values at which REH is rejected for the seven performance metrics

***p* values for tests of REH**		
	Binomial test	**Sum test**
Average Statistical Accuracy	1.81E-07	3.49E-12
Standard Deviation of Statistical Accuracy	4.63E-06	6.76E-10
Maximum Statistical Accuracy	2.35E-04	2.85E-06
Average Combined score	9.82E-07	3.21E-09
Standard Deviation of Combined Score	9.82E-07	3.96E-08
Maximum Combined Score	1.92E-05	5.85E-07
Minimum Combined Score	4.63E-06	6.67E-12

Whichever test we use, the notion that putative differences in expert performance are due to random stressors is overwhelmingly rejected. Table 3 examines the influence of the number of experts and number of calibration variables on the performance metrics.

Table 2: Spearman's Rank Correlation Between Number of Experts and Number of Calibration Variables and Percentile Scores.

	Quantile Avg SA	Quantile STD SA	Quantile Max SA	Quantile Avg. Comb	Quantile STD Comb	Quantile Max Comb
Rank Correlation to # Experts	0.24	0.13	0.03	0.18	0.05	-0.04
Rank Correlation to # Variables	0.40	0.39	0.39	0.31	0.34	0.31

Note. Avg denotes average, SA is for Statistical Accuracy, STD is the standard deviation, Comb. is for combined score, Max is for maximum.

With 49 samples, a rank correlation of 0.24 is significant at the 5% level. As also seen in Table 3, the number of experts is not strongly associated with any of the metrics. The number of calibration variables does appear to exert some influence. Table 3 implies that more calibration variables tend to make the differences between the performance of original experts and randomly scrambled experts greater.

5. Co-benefit of performance weighting for prediction

Statistical accuracy and informativeness are performance metrics for quantifications of uncertainty. There is nothing in these metrics that rewards proximity of the medians to the true values. If these performance metrics enable more accurate predictions of the true values, then this is a collateral benefit, or co-benefit of performance weighting. The (square) Median Deviation for variable i from a distribution with Median_i is defined as follows:

$$\text{MD}_i = (\text{Median}_i - \text{true value}_i)^2.$$

Where true value_i is the true value of the calibration variable i and MD_i is the squared distance between the median elicitation and the truth, true value_i.

MD_i is dependent on the scale of variable i; changing from, say, meters to kilometers will affect the value of MD_i. To aggregate over variables with different scales, the scale dependence must be removed. To compare the proximity of the medians of PW and EW to the realizations, taking the ratio of MD for EW and PW (denoted as EWMD_i and PWMD_i, respectively) per variable removes the scale dependence. These ratios are then aggregated over all variables in a study by taking the geometric mean (geomean):

$$\frac{EWMD}{PWMD} = \left[\prod_{n=1}^{N} \frac{EWMD_i}{PWMD_i}\right]^{1/N}$$ where N is the number of calibration variables.

The geomean is appropriate for aggregating ratios since the geomean of inverse ratios is the inverse of the ratios' geomean and the geomean of ratios is the ratio of geomeans.

While the mean of a linear combination of distributions is the linear combination of their means, the same does not hold for the median[1]. In particular, the median of the equally weighted combination of expert distributions is not equal to the equally weighted combination of their medians. The latter has been presented as "averaging quantiles" as opposed to "averaging probabilities" (Lichtendahl et al., 2013) and has been shown to produce highly overconfident results (Colson & Cooke, 2017). This is so blisteringly obvious that a brief explanation suffices: Consider two experts with 5 and 95 percentiles of [0,1] and [10, 11] respectively. Averaging their percentiles yields a 90% confidence interval of [5, 6]; equally narrow but disjoint from each expert's confidence interval. Experience in expert judgment shows that such situations are not uncommon. "Averaging the probabilities" requires knowledge of the entire distributions. It is more complex but produces distributions more evenly spread over the interval [0, 11]. For purposes of combining distributions, averaging quantiles is a very common and rather severe mistake.

Finding simple point predictions does not require combining distributions. One could also take a simple linear combination of the experts' medians rather than first combining their distributions. This option is worth exploring if only because people will continue combining quantiles in any case. It is useful to assess the loss of performance which this entails.

The linear combinations of medians is denoted Q for quantile aggregation. EWMDQ/PWMDQ is the ratio of square deviations of an equally weighted combination of medians divided by square deviations of a performance weighted combination of medians. This will be compared with EWMDQ/PWiMD, where PWiMD uses the medians of the performance weighted combinations with item specific weights and optimization (Colson & Cooke, 2017). PWiMD represents the "high end" predictor. EWMDQ is the "low end predictor". PWMDQ is easy to compute and hopefully approximates the performance of PWiMD.

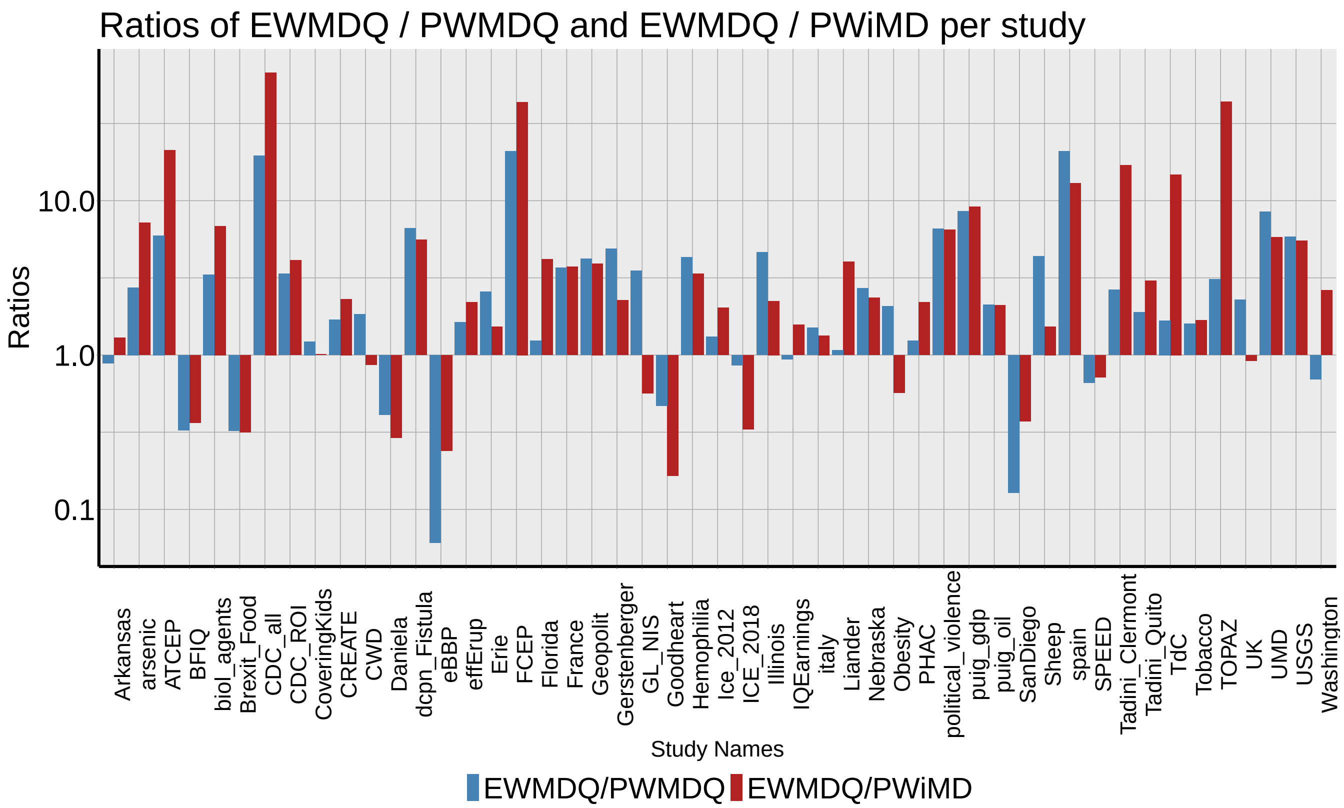
Figure 7 plots EWMDQ/PWMDQ and EWMDQ/PWiMD per study. Values greater than 1 indicate superior predictions relative to EWMDQ. Both PW predictions are superior to EWMDQ, and PWiMD is somewhat better than PWMDQ. Taking the geomeans of these ratios over all studies, EWMDQ/PWMDQ gives 2.05 while EWMDQ/PWiMD gives 2.48. These are ratio's of products of squared distances between the prediction and truth. Taking square roots is equivalent to taking ratios of absolute deviations, which we may characterize as "ratios of distance to the truth". PWMDQ's predictions are in aggregate 43% closer to the truth than predictions of EWMDQ while those of PWiMD are 58% closer. Interestingly, the medians of equally weighted combinations of distributions (not pictured) are 29% closer to the truth, in aggregate, than equally weighted combinations of medians. In other words, if one insists on equal weighting, then it is much better to take the median of an equal weight combination of expert distributions rather than taking an equal weight combination of medians. Using performance weights reduces, but does not eliminate, the advantage of the *median of combinations* as opposed to the com*binations of medians*.

In 37 of the 49 studies the ratios for EWMDQ/PWiMD are greater than 1. The probability of seeing 37 or more ratios greater than 1 if there were really no difference between EW and PW is 2.4E-4. The conclusion is that the predictions based on performance weighted combinations of medians out-

[1] It is easy to see that weighting medians differs from taking the median of weighted combinations of distributions. Consider two lognormal distributions, each with error factor 3 and medians 1 and 9 respectively. An equal weighted combination of their medians is 5, but the median of an equally weighted combination of distributions is 3.

perform equally weighted combinations of medians. Additionally, further improvement is realized by taking the medians of performance combined distributions. The volatility of these ratios per variable is high.

Figure 7: Ratios of EWMDQ / PWMDQ and EWMDQ / PWiMD per study



Note. The ratios are the geomeans per study of ratios of squared distances between the predictions and the true values.

Conclusions

Without denigrating the human capacity for denial, experts do exhibit undeniable differences in their ability to quantify uncertainty. These differences can readily be measured and used to improve performance. The evidence is overwhelming, and it has been overwhelming for some time. However, there are two significant hurdles to applying performance based combinations: (1) Time and effort required for performance measurement and (2) numeracy demands on the analyst.

In most applications, the greatest time and effort is spent in formulating clear questions with operational meaning which address the issues at hand. It is useful to think of expert judgment as a way of obtaining (probabilistic) results from experiments or measurements which are feasible in principle but not in practice. Describing a “thought experiment” is the best way of making absolutely clear what one is asking. These efforts should be made in any case, but the need for operational meaning is easier to ignore if there are no calibration questions. Having formulated questions with clear operational meaning facilitates finding calibration variables from the experts’ field. The impractical experiments or measurements often suggest experiments / measurements which are already performed though not published. Often, as in the recent ‘Ice Sheet 2018’ study, more time is spent agreeing on the best set of calibration variables than in generating them. That said, finding good calibration variables does require a deep dive in the subject matter, which is greatly aided by having a domain expert on the analysis team.

One-on-one interviews cost time and money, although good online meeting tools bring these costs way down. One-on-one elicitation enables the analysts to better plumb experts' reasoning. Supervised plenary elicitation in which experts meet, discuss, and then individually perform the elicitation offer advantages of speed and disadvantages in loss of individual engagement.

Sending a questionnaire in the mail to a large set of experts in the hope that a fair number will respond is discouraged for purposes of uncertainty quantification. Expert surveys should be sharply distinguished from structured expert judgment.

Despite all this, the most difficult hurdle is the second: finding qualified analysts. Mathematicians, statisticians, engineers and scientists know that the Classical Model is not a heavy lift[2]. Many have conducted successful studies in their chosen fields. The analyst must be able to explain the method to the experts and to the problem owners, so that they in turn can explain it up the chain. If a problem owner is unable to explain the method to his/her superiors, (s)he is unlikely to adopt it. The analyst must be comfortable with certain relevant concepts such as statistical likelihood, *p* values, Shannon information, scoring rules, distributions, densities, quantiles, etc. The analyst must be able to explain why (s)he is doing things this way and not using any of the slap dash approaches proliferating the blogosphere. Some knowledge of foundations is needed to explain why uncertainty is represented as subjective probability and not as fuzziness, imprecision, degree of possibility, certainty factors, to name a few. Writing up the results in a clear an accurate fashion requires more than a nodding acquaintance with all these concepts.

[2] Cooke (2015) esp. the supplementary online information is written to bring neophytes up to speed. The TU Delft will launch a free online course on expert judgment in October 2019.

References

Bamber, J. L., Oppenheimer, Kopp, R. E., Aspinall, W.P., Cooke, Roger M., (2019) Ice sheet contributions to future sea level rise from structured expert judgement, accepted for publication in PNAS.

Colson .A., and Cooke, R.M., (2018) Expert Elicitation: Using the Classical Model to Validate Experts' Judgments. Review of Environmental Economics and Policy, Volume 12, Issue 1, 1 February 2018, Pages 113–132, https://doi.org/10.1093/reep/rex022 https://academic.oup.com/reep/article/12/1/113/4835830

Colson, A. and Cooke, R.M., (2017) Cross Validation for the Classical Model of Structured Expert Judgment, Reliability Engineering and System Safety, Volume 163, July 2017, Pages 109–120 http://dx.doi.org/10.1016/j.ress.2017.02.003

Colson, A. Cooke, R.M., Lutter, Randall, (2016) How Does Breastfeeding Affect IQ? Applying the Classical Model of Structured Expert Judgment, Resources for the Future, RFF DP16-28 http://www.rff.org/research/publications/how-does-breastfeeding-affect-iq-applying-classical-model-structured-expert

Colson, A.R., Megiddo, I., Alvarez-Uria, G., Gandra, S., Bedford, T. Morton, A., Cooke, R.M., Laxminarayan, R., (2019) "Quantifying Uncertainty about Future Antimicrobial Resistance: Comparing Structured Expert Judgment and Statistical Forecasting Methods", almost accepted at PLoS One.

Colson, A.R., Megiddo, I., Alvarez-Uria, G., Gandra, S., Bedford, T. Morton, A., Cooke, R.M., Laxminarayan, R., (2019) "Quantifying Uncertainty about Future Antimicrobial Resistance: Comparing Structured Expert Judgment and Statistical Forecasting Methods", accepted at PLoS One.

Cooke Roger M.,(1991) Experts in Uncertainty; Opinion and Subjective Probability in Science, Oxford University Press; New York Oxford, 321 pages.; ISBN 0-19-506465-8

Cooke, Roger M., A Theory of Weights for Combining Expert Opinions, Report 87-25. Dept. of Mathematics, Delft University of Technology, 1987.

Cooke, Roger M., Goossens, L.H.J. (2008) Special issue on expert judgment Reliability Engineering & System Safety, 93, 657-674, Available online 12 March 2007, Issue 5, May 2008.

Cooke, Roger M., Mendel, M., Thijs, W.,(1988) "Calibration and Information in Expert Resolution". Automatica, 24, 1, 87-94, 1988.

Cooke, Roger M. (2015) "Messaging climate change uncertainty with Supplementary Online Material" Nature Climate Change 5, 8–10 (2015) doi:10.1038/nclimate2466 Published online 18 December 2014 http://www.nature.com/nclimate/journal/v5/n1/full/nclimate2466.html

Eggstaff,J.W., Mazzuchi,T.A. Sarkani, S. (2014) The Effect of the Number of Seed Variables on the Performance of Cooke's Classical Model, Reliability Engineering and System Safety 121 (2014) 72–82. DOI: 10.1016/j.ress.2013.07.015

Ismail, Raveem and Ried, Scott (2015) "Ask the Experts" , The Actuary, the official magazine of the Institute and Faculty of Actuaries 6/15/2016.

Lichtendahl Jr, K.C., Grushka-Cockayne, Y., Winkler, R.L., (2013) Is it better to average probabilities or quantiles? Management Science 59 (7), 1594-1611

Lutter, R. Colson, A. and Cooke, R.M. (2017) Effects of Increases in IQ in India on the Present Value of Lifetime Earnings A Structured Expert Judgment Study, RFF WP 17-18 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3094387

Marti, H. D., Mazzuchi, T.A., and Cooke R.M. (2019) Are Performance Weights Beneficial? Investigating the Random Expert Hypothesis" appearing in Hanea, Nane, French and Bedford.

Puig, Daniel, Morales-Nápoles, Oswaldo, Bakhtiari, Fatemeh, Landa, Gissela (2018), The accountability imperative for quantifying the uncertainty of emission forecasts: evidence from Mexico, Climate Policy, doi: 10.1080/14693062.2017.1373623, 742-751

Tadini, A., M. Bisson, A. Neri, R. Cioni, A. Bevilacqua and W. P. Aspinall (2017), Assessing future vent opening locations at the Somma‐Vesuvio volcanic complex: 1. A new information geodatabase with uncertainty characterizations, J. Geophys. Res. Solid Earth, 122, doi:10.1002/2016JB013858.

Winkler, R.L. Grushka-Cockayne, Y., Lichtendahl Jr. K.C., and Jose, V.R.R. (2018) , Averaging Probability Forecasts: Back to the Future, Working Paper | HBS Working Paper Series | 2018

Appendix A

Data references table

Study Name	Reference
UMD	Koch, Benjamin J., Filoso, S., Cooke, R. M. Hosen, J. D., Colson, A.R. Febria, Catherine M., Palmer, M. A. , (2015) Nitrogen in stormwater runoff from Coastal Plain watersheds: The need for empirical data, reply to Walsh , Elementa DOI 10.12952/journal.elementa.000079. https://www.elementascience.org/articles/79 Koch, Benjamin J., Febria, Catherine M. , Cooke, Roger M. Hosen, Jacob D. , Baker, Matthew E. , Colson, Abigail R. Filoso, Solange, Hayhoe, Katharine, Loperfido, J.V. , Stoner, Anne M.K. , Palmer, Margaret A. , (2015) Suburban watershed nitrogen retention: Estimating the effectiveness of storm water management structures, Elementa, DOI 10.12952/journal.elementa.000063 https://www.elementascience.org/articles/63
USGS	Newhall, C. G., & Pallister, J. S. (2015). Using multiple data sets to populate probabilistic volcanic event trees. In Volcanic Hazards, Risks and Disasters (pp. 203-232).
arsenic	Hanzich, J.M. (2007) Achieving Consensus: An Analysis Of Methods To Synthesize Epidemiological Data For Use In Law And Policy. Department of Public Health & Primary Care, Institute Of Public Health, University of Cambridge; unpublished MPhil thesis, 66pp + appendices.
Biol Agents	Aspinall & Associates (2006). REBA Elicitation. Commercial-in-confidence report, 26pp.
Geopolit	Ismail and Reid (2006). “Ask the Experts” presentation
ATCEP	Morales-Nápoles, O., Kurowicka, D., & Cooke, R. (2008). EEMCS final report for the causal modeling for air transport safety (CATS) project.
Daniela	Forys, M.B., Kurowicka, D., Peppelman, B.(2013) "A probabilistic model for a gas explosion due to leakages in the grey cast iron gas mains" Reliability Engineering & System Safety volume 119, issue , year 2013, pp. 270 - 279.
eBBP	Tyshenko, M.G., S. ElSaadany, T. Oraby, M. Laderoute, J. Wu, W. Aspinall and D. Krewski (2011) Risk Assessment and Management of Emerging Blood-Borne Pathogens in Canada: Xenotropic Murine Leukaemia Virus-Related Virus as a Case Study for the Use of a Precautionary Approach. Chapter in: *Risk Assessment* (ISBN 979-953-307-765-8). Cashman, N.R., Cheung, R., Aspinall, W., Wong, M. and Krewski, D. (2014) Expert Elicitation for the Judgment of Prion Disease Risk Uncertainties associated with Urine-derived and Recombinant Fertility Drugs. Submitted to: Journal of Toxicology and Environmental Health
create	Bier V.M, Kosanoglu, F, Shin J, unpublished data, nd.
effErupt	Aspinall, W.P. (2012) Comment on “Social studies of volcanology: knowledge generation and expert advice on active volcanoes” by Amy Donovan, Clive Oppenheimer and Michael Bravo [*Bull Volcanol* (2012) 74:677-689] Bulletin of Volcanology, 74, 1569-1570. doi: 10.1007/s00445-012-0625-x
erie	Colson, Abigail R., Sweta Adhikari, Ambereen Sleemi, and Ramanan Laxminarayan. (2015) “Quantifying Uncertainty in Intervention Effectiveness with Structured Expert Judgment: An Application to Obstetric Fistula.” BMJ Open, 1–8. doi:10.1136/bmjopen-2014-007233.

	Cooke, R.M., Wittmann, M.E., Lodge, D.M., Rothlisberger, J.D., Rutherford E.S., Zhang, H. and Mason, D.M. (2014) "Out-of-Sample Validation for Structured Expert Judgment of Asian Carp Establishment in Lake Erie", Integrated Environmental Assessment and Management, open access. DOI: 10.1002/ieam.1559 Zhang, H, Rutherford E.S., Mason, D.M., Breck, J,T,, Wittmann M.E., Cooke R.M., Lodge D.M., Rothlisberger J.D., Zhu X., and Johnson, T B., (2015) Forecasting the Impacts of Silver and Bighead Carp on the Lake Erie Food Web, Transactions of the American Fisheries Society, Volume 145, Issue 1, pp 136-162, DOI:10.1080/00028487.2015.1069211
FCEP	Leontaris, G., & Morales-Nápoles, O. (2018). ANDURIL—A MATLAB toolbox for ANalysis and Decisions with UnceRtaInty: Learning from expert judgments. SoftwareX, 7, 313-317.
Sheep	Hincks, T., Aspinall, W. and Stone, J. (2015) Expert judgement elicitation exercise to evaluate Sheep Scab control measures: Results of the Bayesian Belief Network analysis. University of Bristol PURE Repository Working Paper (forthcoming).
hemophilia	Fischer K, Lewandowski D, Janssen MP. Estimating unknown parameters in haemophilia using expert judgement elicitation. Haemophilia. 2013 Sep;19(5):e282-e288.
Liander	Forys, M.B., Kurowicka, D., Peppelman, B.(2013) "A probabilistic model for a gas explosion due to leakages in the grey cast iron gas mains" Reliability Engineering & System Safety volume 119, issue , year 2013, pp. 270 - 279.
PHAC	Oraby,T., Tyshenko, M.G., Westphal, M., Darshan, S., Croteau, M., Aspinall, W., Elsaadany, S., Cashman, N. and Krewski, D. (2011) Using Expert Judgments to Improve Chronic Wasting Disease Risk Management in Canada. Journal of Toxicology and Environmental Health, in press. Volume 74, Issue 2-4, 2011 Special Issue: Prion Research in Perspective 2010
TOPAZ	Scourse, E., Aspinall, W.P. and Chapman, N. (2014) Using expert elicitation to characterise long-term tectonic risks to radioactive waste repositories in Japan. Journal of Risk Research, doi: 10.1080/13669877.2014.971334
SPEED	Hicks, A., Barclay, J., Simmons, P. and Loughlin, S. (2014). "An interdisciplinary approach to volcanic risk reduction under conditions of uncertainty: a case study of Tristan da Cunha." Nat. Hazards Earth Syst. Sci. 14(7): 1871-1887. Doi: 10.5194/nhess-14-1871-2014. www.nat-hazards-earth-syst-sci-discuss.net/1/7779/2013/ Bevilacqua, A., Isaia, R., Neri, A., Vitale, S., Aspinall, W.P. and eight others (2015) Quantifying volcanic hazard at Campi Flegrei caldera (Italy) with uncertainty assessment: I. Vent opening maps. Journal of Geophysical Research - Solid Earth; AGU. doi:10.1002/2014JB011775
TDC	Scourse, E., Aspinall, W.P. and Chapman, N. (2014) Using expert elicitation to characterise long-term tectonic risks to radioactive waste repositories in Japan. Journal of Risk Research, doi: 10.1080/13669877.2014.971334
GL	Rothlisberger,J.D. Finnoff, D.C. Cooke,R.M. and Lodge, D.M. (2012) "Ship-borne nonindigenous species diminish Great Lakes ecosystem services" Ecosystems (2012) 15: 462–476 DOI: 10.1007/s10021-012-9522-6 Rothlisberger, J.D., Lodge, D.M. Cooke, R.M. and Finnoff, D.C. (2009) "Future declines of the binational Laurentian Great Lakes fisheries: recognizing the importance of environmental and cultural change" *Frontiers in Ecology and the Environment;* doi:10.1890/090002
Goodheart	Goodheart, B. (2013). Identification of causal paths and prediction of runway incursion risk by means of Bayesian belief networks. Transportation Research Record: Journal of the Transportation Research Board, (2400), 9-20.

Ice	Bamber, J.L., and Aspinall, W.P., (2012) An expert judgement assessment of future sea 1evel rise from the ice sheets, Nature Climate Change, PUBLISHED ONLINE: January 6, 2012 | DOI: 10.1038/NCLIMATE1778. http://www.nature.com/nclimate/journal/vaop/ncurrent/full/nclimate1778.html
puig-gdp	Puig, D., Morales-Nápoles, O., Bakhtiari, F., & Landa, G. (2018). The accountability imperative for quantifying the uncertainty of emission forecasts: evidence from Mexico. Climate Policy, 18(6), 742-751.
puig-oil	Puig, D., Morales-Nápoles, O., Bakhtiari, F., & Landa, G. (2018). The accountability imperative for quantifying the uncertainty of emission forecasts: evidence from Mexico. Climate Policy, 18(6), 742-751.
YTBID (CDC)	Colson, Abigail R., M.A. Cohen, S. Regmi, A. Nandi, R. Laxminarayan (2015) "Structured Expert Judgment for Informing the Return on Investment in Surveillance: The Case of Environmental Public Health Tracking." Working Paper. Center for Disease Dynamics, Economics & Policy.
Gerestenberger	Gerstenberger, M. C., et al. (2016). "A Hybrid Time-Dependent Probabilistic Seismic-Hazard Model for Canterbury, New Zealand." Seismological Research Letters. Vol. 87 Doi: 10.1785/0220160084 Gerstenberger, M.C.; McVerry, G.H.; Rhoades, D.A.; Stirling, M.W. (2014) Seismic hazard modeling for the recovery of Christchurch, New Zealand.*Earthquake Spectra, 30(1):* 17-29; doi: 10.1193/021913EQS037M Gerstenberger, M.C.; Christophersen, A.; Buxton, R.; Allinson, G.; Hou, W.; Leamon, G.; Nicol, A. (2013) Integrated risk assessment for CCS. p. 2775-2782; doi: 10.1016/j.egypro.2013.06.162 IN: Dixon, T.; Yamaji, K. (eds) 11th International Conference on Greenhouse Gas Control Technologies, 18th-22nd November 2012, Kyoto International Conference Center, Japan. Elsevier. Energy procedia 37
CWD	Tyshenko, M.G., ElSaadany, S., Oraby, T., Darshan, S., Catford, A., Aspinall, W., Cooke, R. and Krewski, D. (2012) Expert judgement and re-elicitation for prion disease risk uncertainties. *International Journal of Risk Assessment and Management*, 16(1-3), 48-77. doi:10.1504/IJRAM.2012.047552 Tyshenko, M.G., S. ElSaadany, T. Oraby, S. Darshan, W. Aspinall, R. Cooke, A. Catford, and D. Krewski (2011) Expert elicitation for the judgment of prion disease risk uncertainties. *J Toxicol Environ Health* A.; 74(2-4):261-285. Oraby,T., Tyshenko, M.G., Westphal, M., Darshan, S., Croteau, M., Aspinall, W., Elsaadany, S., Cashman, N. and Krewski, D. (2011) Using Expert Judgments to Improve Chronic Wasting Disease Risk Management in Canada. Journal of Toxicology and Environmental Health, in press. Volume 74, Issue 2-4, 2011 Special Issue: Prion Research in Perspective 2010
Nebraska	Attribution study for Robert Wood Johnson Covering Kids & Families in Pennsylvania, Washington, Nebraska, Illinois, Arkansas, and Florida, conducted by Center for Disease Dynamics, Economics & Policy, 2012.
SanDiego	Attribution study for Robert Wood Johnson Covering Kids & Families in Pennsylvania, Washington, Nebraska, Illinois, Arkansas, and Florida, conducted by Center for Disease Dynamics, Economics & Policy, 2012.
BFIQ	Colson, A. Cooke, R.M., Lutter, Randall, (2016) How Does Breastfeeding Affect IQ? Applying the Classical Model of Structured Expert Judgment, Resources for the Future, RFF DP16-28

	http://www.rff.org/research/publications/how-does-breastfeeding-affect-iq-applying-classical-model-structured-expert
France	Abigail R. Colson, Itamar Megiddo, Gerardo Alvarez-Uria, Sumanth Gandra, Tim Bedford, Alec Morton, Roger M. Cooke , Ramanan Laxminarayan (ns). "Quantifying Uncertainty about Future Antimicrobial Resistance: Comparing Structured Expert Judgment and Statistical Forecasting Methods."
Italy	Abigail R. Colson, Itamar Megiddo, Gerardo Alvarez-Uria, Sumanth Gandra, Tim Bedford, Alec Morton, Roger M. Cooke , Ramanan Laxminarayan (ns). "Quantifying Uncertainty about Future Antimicrobial Resistance: Comparing Structured Expert Judgment and Statistical Forecasting Methods."
Spain	Abigail R. Colson, Itamar Megiddo, Gerardo Alvarez-Uria, Sumanth Gandra, Tim Bedford, Alec Morton, Roger M. Cooke , Ramanan Laxminarayan (ns). "Quantifying Uncertainty about Future Antimicrobial Resistance: Comparing Structured Expert Judgment and Statistical Forecasting Methods."
UK	Abigail R. Colson, Itamar Megiddo, Gerardo Alvarez-Uria, Sumanth Gandra, Tim Bedford, Alec Morton, Roger M. Cooke , Ramanan Laxminarayan (ns). "Quantifying Uncertainty about Future Antimicrobial Resistance: Comparing Structured Expert Judgment and Statistical Forecasting Methods."
Arkansas	Attribution study for Robert Wood Johnson Covering Kids & Families in Pennsylvania, Washington, Nebraska, Illinois, Arkansas, and Florida, conducted by Center for Disease Dynamics, Economics & Policy, 2012.
CoveringKids	Attribution study for Robert Wood Johnson Covering Kids & Families in Pennsylvania, Washington, Nebraska, Illinois, Arkansas, and Florida, conducted by Center for Disease Dynamics, Economics & Policy, 2012.
dcpn_Fistula	Aspinall,W. Devleesschauwer, B. Cooke, R.M., Corrigan,T., Havelaar, A.H., Gibb, H., Torgerson, P., Kirk, M., Angulo, F., Lake, R., Speybroeck, N., and Hoffmann, S. (2015) World Health Organization estimates of the relative contributions of food to the burden of disease due to selected foodborne hazards: a structured expert elicitation. PLOS ONE, : January 19, 2016 DOI: 10.1371/journal.pone.0145839.
Florida	Attribution study for Robert Wood Johnson Covering Kids & Families in Pennsylvania, Washington, Nebraska, Illinois, Arkansas, and Florida, conducted by Center for Disease Dynamics, Economics & Policy, 2012.
Illinois	Attribution study for Robert Wood Johnson Covering Kids & Families in Pennsylvania, Washington, Nebraska, Illinois, Arkansas, and Florida, conducted by Center for Disease Dynamics, Economics & Policy, 2012.
Obesity	Colson, Abigail R., R.M. Cooke, R. Laxminarayan. (2015) "Attributing Impact to a Charitable Foundation's Programs with Structured Expert Judgment." Working Paper. Center for Disease Dynamics, Economics & Policy.
Tobacco	Colson, Abigail R., R.M. Cooke, R. Laxminarayan. (2015) "Attributing Impact to a Charitable Foundation's Programs with Structured Expert Judgment." Working Paper. Center for Disease Dynamics, Economics & Policy.

Washington	Attribution study for Robert Wood Johnson Covering Kids & Families in Pennsylvania, Washington, Nebraska, Illinois, Arkansas, and Florida, conducted by Center for Disease Dynamics, Economics & Policy, 2012.
cdc-roi	Colson, Abigail R., M.A. Cohen, S. Regmi, A. Nandi, R. Laxminarayan (2015) “Structured Expert Judgment for Informing the Return on Investment in Surveillance: The Case of Environmental Public Health Tracking.” Working Paper. Center for Disease Dynamics, Economics & Policy.
IQ-earn	Randall Lutter, Abigail Colson, and Roger Cooke (ns), (ns), “Effects of Increases in IQ in India on the Present Value of Lifetime Earnings
Tadini_Quito	Tadini, A., M. Bisson, A. Neri, R. Cioni, A. Bevilacqua and W. P. Aspinall (2017), Assessing future vent opening locations at the Somma ‐ Vesuvio volcanic complex: 1. A new information geodatabase with uncertainty characterizations, J. Geophys. Res. Solid Earth, 122, doi:10.1002/2016JB013858.
Tadini_Clermont	Tadini, A., M. Bisson, A. Neri, R. Cioni, A. Bevilacqua and W. P. Aspinall (2017), Assessing future vent opening locations at the Somma ‐ Vesuvio volcanic complex: 1. A new information geodatabase with uncertainty characterizations, J. Geophys. Res. Solid Earth, 122, doi:10.1002/2016JB013858.
PoliticalViolence	In preparation
Brexit_food	In preparation
ICE_2018	Bamber, J. L., Oppenheimer, Kopp, R. E., Aspinall, W.P., Cooke, Roger M., (2019) Ice sheet contributions to future sea level rise from structured expert judgement, accepted for publication in PNAS.