Appendix 1. Classical Model Performance Measures and Combination

based on courses given for NASA, FAA and various companies and labs from Cooke, R.M. *Experts in Uncertainty*, Oxford University Press, 1991

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

- **REEP** Validation for Classical Model
- **RESS** Cross Validation extensive SOM on validation, aggregation
- **Comp & OR** Quantifying Info Security Risks
- **PNAS** SEJ ice sheets SOM SEJ data and elicitation protocol
- **PLoS 1** Evaluation of Performance: WHO SEJ study of global burden of disease
- **Elementa** Stormwater Management in Chesapeake Bay; extensive SOM elicitation protocol
- **Cons Bio** Impacts of Asian Carp Invasion Lake Erie - SEJ SOM
- **RFF** SEJ Breast Feeding and IQ & AStA
- **RFF** SEJ IQ and Earnings
- **Nature Climate Change:** Messaging climate Uncertainty with extensive **SOM** on representation of uncertainty

Videos

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

Online Course

SEJ TU Delft

There are two generic, quantitative measures of performance, *statistical accuracy* (aka *calibration)* and *information*. Loosely, statistical accuracy denotes the statistical likelihood that a set of experimental results correspond, in a statistical sense, with an expert's assessments. Information measures the degree to which a distribution is concentrated. To simplify the exposition we assume that the 5%, 50% and 95% values were elicited.

Statistical accuracy

Assume that each expert assess 5, 50 and 95 percentiles for each quantity. Each expert divides the range of possible values into 4 inter quantile intervals for which his/her probabilities are known, namely $p_1 =$ 0.05: less than or equal to the 5% value, $p_2 = 0.45$: greater than the 5% value and less than or equal to the 50% value, etc.

If N quantities are assessed, each expert may be regarded as a statistical hypothesis, namely that each realization falls in one of the four inter-quantile intervals with probability vector

$p = (0.05, 0.45, 0.45, 0.05)$.

Suppose we have realizations $x_1,...x_N$ of these quantities. We may then form the sample distribution of the expert's inter quantile intervals as:

$s_1(e) = \#\{i \mid x_i \leq 5\% \text{ quantile}\}/N$
$s_2(e) = \#\{i \mid 5\% \text{ quantile} < x_i \leq 50\% \text{ quantile}\}/N$
$s_3(e) = \#\{i \mid 50\% \text{ quantile} < x_i \leq 95\% \text{ quantile}\}/N$
$s_4(e) = \#\{i \mid 95\% \text{ quantile} < x_i \}/N$
$s(e) = (s_1(e), \ldots s_4(e))$

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

$$2NI(s(e)| p) = 2N\ \Sigma_{i=1..4}\, s_i\, \ln(s_i / p_i) \qquad \text{(A1.1)}$$

is asymptotically distributed as a chi-square variable with 3 degrees of freedom. This is the so-called likelihood ratio statistic, and $I(s \mid p)$ is the relative information of distribution *s* with respect to *p*. If we extract the leading term of the logarithm we obtain the familiar chi-square test statistic for goodness of fit. There are advantages in using the form in (A1.1) (Cooke 1991).

If after a few realizations the expert were to see that all realization fell outside his 90% central confidence intervals, he might conclude that these intervals were too narrow and might broaden them on subsequent assessments. This means that for this expert the uncertainty distributions are *not* independent, and he learns from the realizations. Expert learning is not a goal of an expert judgment study and his joint distribution is not elicited. Rather, the decision maker wants experts who do not need to learn from the elicitation. Hence the decision maker scores expert *e* as the statistical likelihood of the hypothesis

H_e: *"the inter quantile interval containing the true value for each variable is drawn independently from probability vector p."*

A simple test for this hypothesis uses the test statistic *(A1.1)*, and the likelihood, or p-value, or **statistical accuracy score** of this hypothesis, is:

$$Sa(e) = p\text{-value} = Prob\{2NI(s(e)| p) \geqslant r \mid H_e\}$$

where *r* is the value of *(A1.1)* 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. Statistical accuracy 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 *(A1.1)*, without changing the sample distribution *s*. In some cases involving multiple realizations of one and the same assessment, the effective number of seed variables is based on the number of assessments and not the number of realizations.

Although the statistical accuracy score uses 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 accurate. Low scores, near zero, mean that it is unlikely that the expert's probabilities are correct.

Information
The second scoring variable is information. Loosely, the information in a distribution is the degree to which the distribution is concentrated. Information cannot be measured absolutely, but only with respect to a background measure. Being concentrated or "spread out" is measured relative to some other distribution.

Measuring information requires associating a density to each variable based on each expert's quantile assessments. To do this, we use the unique density that complies with the experts' quantiles and is minimally informative with respect to a background measure. This density can easily be found with the method of Lagrange multipliers. 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 indeed it must be the same for all experts; instead it is chosen by the analyst.

The uniform and log-uniform background measures require an *intrinsic range* on which these measures are concentrated. The classical model 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

$$I^* = [L^*, U^*]; L^* = L - k(U\text{-}L)/100; \;\; U^* = U + k(U\text{-}L)/100.$$

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 expert e on assessments for uncertain quantities 1…N is

$$Inf(e) = \text{Average Relative information wrt Background} = (1/N) \;\; \Sigma_{i=1..N} I(f_{e,i} \mid g_i)$$

where g_i is the background density for variable i and $f_{e,i}$ is expert e's density for item i. This is proportional to the relative information of the expert's joint distribution given the background, under the assumption that the variables are independent. As with statistical accuracy, the assumption of independence here reflects a desideratum of the decision maker and not an elicited feature of the expert's joint distribution. The information score does not depend on the realizations. An expert can give himself a high information score by choosing his quantiles very close together.

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

Of course, other measures of concentrated-ness could be contemplated. The above information score is chosen because it is

- familiar
- tail insensitive
- scale invariant
- slow

The latter property means that relative information is a slow function; large changes in the expert assessments produce only modest changes in the information score. This contrasts with the likelihood function in the statistical accuracy score, which is a very fast function. This causes the product of statistical accuracy and information to be driven by the statistical accuracy score. It also means that modest changes in informativeness correspond to sizeable changes in the distributions. Increasing informativeness by a factor *2* roughly corresponds to halving the distance between the 95 and 5 percentiles.

Combination: Decision Maker

The **combined score** of expert e will serve as an (unnormalized) weight for e:

$$w_\alpha(e) = Sa\,(e) \times Inf\,(e) \times 1_\alpha(Sa(e) \geq \alpha), \qquad (A1.2)$$

where $1_\alpha(Sa\,(e)) = 1$ if Sa$(e) \geq \alpha$, and is zero otherwise. The combined score thus depends on α. If Sa(e) falls below cut-off level α expert e is unweighted. The presence of a cut-off level is imposed by the requirement that the combined score be an asymptotically strictly proper scoring rule. That is, an expert maximizes his/her long run expected score by and only by ensuring that his probabilities $p= (0.05, 0.45, 0.45, 0.05)$ correspond to his/her true beliefs. α is similar to a significance level in simple hypothesis testing, but its origin is indeed different. The goal of scoring is not to “reject” hypotheses, but to measure “goodness” with a strictly proper scoring rule.

A combination of expert assessments is called a "decision maker" (DM). All decision makers discussed here are examples of linear pooling. The classical model is essentially a method for deriving weights in a linear pool. "Good expertise" corresponds to good statistical accuracy (high statistical likelihood, high p-value) and high information. We want weights which reward good expertise and which pass these virtues on to the decision maker.

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 product of statistical accuracy and information. Solving this problem on 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 decision maker 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 strictly scoring rule constraint on the weighing scheme.

The scoring rule constraint requires the term 1_α*(statistical accuracy score),* but does not say what value of α we should choose. Therefore, we choose α so as to maximize the combined score of the resulting decision maker. Let $DM_\alpha(i)$ be the result of linear pooling for item *i* with weights proportional to (*A1*.2):

$$DM_\alpha(i) = \Sigma_{e=1,..E}\, w_\alpha(e)\, f_{e,i} \;/\; \Sigma_{e=1,..E}\, w_\alpha(e) \qquad (A1.3)$$

The *optimized global weight DM* is DM_{α^*} where α^* maximizes

$$statistical\ accuracy\ score(DM_\alpha) \times information\ score(DM_\alpha). \qquad (A1.4)$$

This weight is termed global because the information score is based on all the assessed seed 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:

$$w_\alpha(e,i) = 1_\alpha(statistical\ accuracy\ score) \times statistical\ accuracy\ score(e) \times I(f_{e,i} \mid g_i) \qquad (A1.5)$$

For each α we define the Item weight DM_α for item *i* as

$$IDM_\alpha(i) = \Sigma_{e=1,..E}\, w_\alpha(e,i)\, f_{e,i} \;/\; \Sigma_{e=1,..E}\, w_\alpha(e,i) \qquad (A1.6)$$

The *optimized item weight DM* is IDM_{α^*} where α^* maximizes

$$statistical\ accuracy\ score(IDM_\alpha) \times information\ score(IDM_\alpha). \qquad (A1.7)$$

The non-optimized versions of the global and item weight DM’s are obtained simply by setting $\alpha = 0$.

Item weights are potentially more attractive as they allow an expert to up- or down- weight him/herself for individual items according to how much (s)he feels (s)he knows about that item. "knowing less" means choosing quantiles further apart and lowering the information score for that item. Of course, 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. Both item and global weights can be pithily described as optimal weights under a strictly proper scoring rule constraint. In both global and item weights statistical accuracy dominates over information, information serves to modulate between more or less equally well calibrated experts. Definitions and proofs of these scoring rule properties are found in Cooke, R.M. *Experts in Uncertainty*, Oxford University Press, 1991.

Since any combination of expert distributions yields assessments for the seed variables, any combination can be evaluated on the seed variables. In particular, we can compute the statistical accuracy and the information of any proposed decision maker. We should hope that the decision maker would perform better than the result of simple averaging of distributions, called the *equal weight DM*, and we should also hope that the proposed DM is not worse than the best expert in the panel. The global and item weight DM's discussed above (optimized or not) are *Performance based DM's.* In general the optimized global weight DM is used, unless the optimized item weight DM is markedly superior.

The optimization in *(A1.5)* and *(A1.7)* often causes experts to be unweighted, even experts 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 loss of experts.

APPENDIX 2. In Sample Validation

The histogram of calibration variables and the graph of *p*-value scores (statistical accuracy scores) are given in Figure A2.1.

Figure A2.1*: Calibration frequencies (left) for all 526 post-2006 experts who assessed all calibration variables and p-value scores (right) for all 530 experts not accounting for different numbers of assessed calibration variables. This includes 4 experts who skipped some calibration variables. The traditional 5% threshold for simple hypothesis testing is given as a red line.*

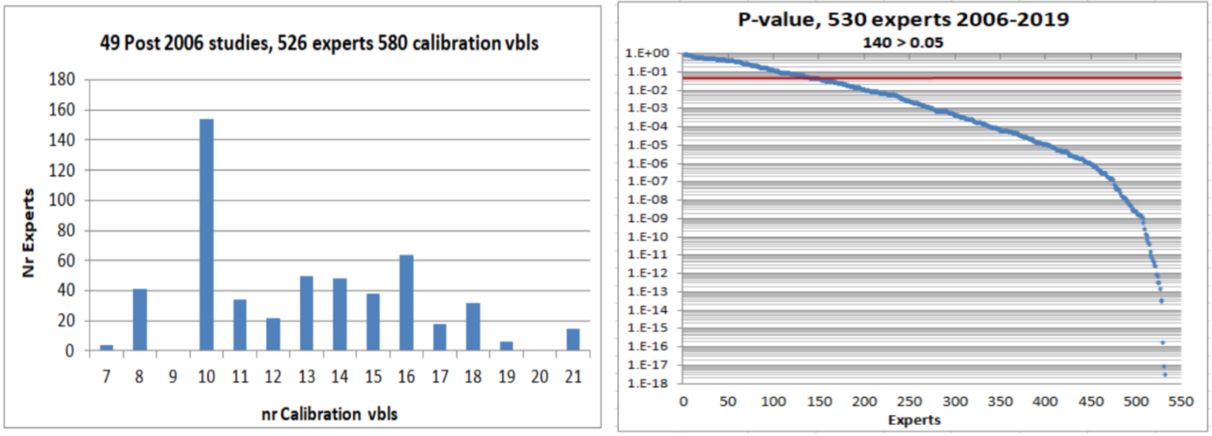
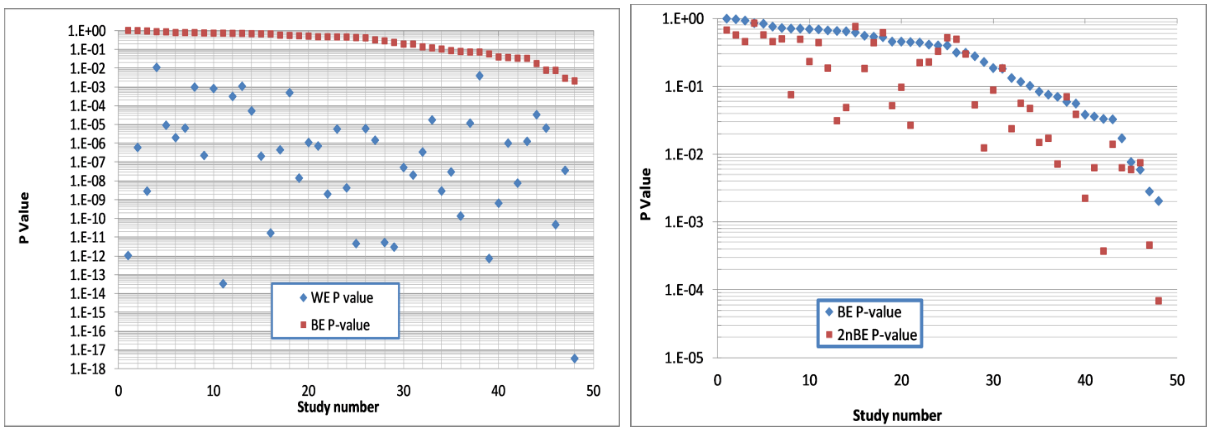


Figure A2.2 shows the best and worst *p-values* per study for all *49* post-2006 studies (left) and the two best performing experts (right). There are generally *4* or more orders of magnitude in statistical accuracy scores between the best and worst expert per study. Despite the fact that only *140* of the *530* 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.

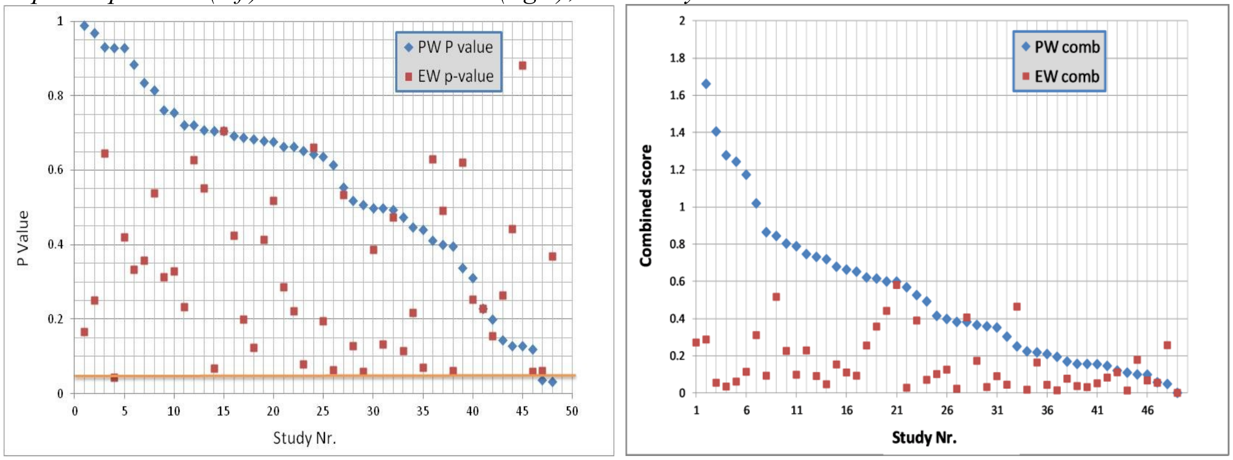
Figure A2.2: Best and worst expert (BE and WE, respectively) p-values or Statistical Accuracy (left) and p-values of best two experts in terms unnormalized weight (combined score of statistical accuracy × informativeness) (right), per study, 2006-2019 data.



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 A2.3 shows the in-sample results for the 2006-2019 data. The combined score is the product of the (dimensionless) statistical accuracy and informativeness scores. 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 A2.3: 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.



Appendix 3 Data and Data References

Post-2006 Expert Data

In addition to the *33* post-2006 studies studied in (Colson and Cooke 2017; 2018), 16 post 2016 studies have been added, as summarized in Table A3.1.

Table A3. 1 Expert judgment studies are shown with the number of calibration variables and experts, post-2006 (post-2016 bolded), asterisks denote studies where variables were excluded in section 5 owing to equality of prediction and realization. Four experts who did not assess all calibration variables are excluded.

Study	# Expert	# Calib Vbls	Subject
Arkansas	4	10	Grant effectiveness, child health insurance enrollment

arsenic	9	10	Air quality levels for arsenic
ATCEP	5	10	Air traffic controllers human Error
BFIQ	7	11	**Breastfeeding and IQ**
biol_agents*	12	12	Human dose-response curves for bioterror agents
Brexit_Food	10	10	**Food price change after Brexit**
CDC_all	48	14	**Global burden of disease**
CDC_ROI	20	10	Return on investment for CDC warnings
CoveringKids	5	10	Grant effectiveness, child health insurance enrollment
CREATE	7	10	Terrorism
CWD	14	10	Transmission risks: chronic wasting disease from deer to humans
Daniela	4	7	Fire prevention and control
dcpn_Fistula	8	10	Effectiveness of obstetric fistula repair
eBBP**	14	15	XMRV blood/tissue infection transmission risks
effErup	14	8	Icelandic fissure eruptions: source characterization
Erie	10	15	Establishment of Asian Carp in Lake Erie
FCEP	5	8	Flight crew human rror
Florida	7	10	Grant effectiveness, child health insurance enrollment
France	5	10	**Future antimicrobial eesistance in France**
Geopolit	9	16	**Geopolitics**
Gerstenberger	12	13	Probabilistic Seismic-Hazard Model for Canterbury
GL_NIS	9	13	Costs of invasive species in Great Lakes
Goodheart	6	10	Airport safety
Hemophilia	18	8	Hemophilia
Ice_2012	10	11	Sea level rise from Ice Sheets melting due to global warming
ICE_2018	20	16	**Future see level rise**
Illinois	5	10	Grant effectiveness, child health insurance enrollment
IQEarnings	8	11	**Effect IQ in India on present value of lifetime earnings**
Italy	4	8	**Future antimicrobial eesistance in Italy**
Liander	11	10	Underground cast iron gas-lines
Nebraska	4	10	Grant effectiveness, child health insurance enrollment
Obesity	4	10	Grant effectiveness, childhood obesity
PHAC	10	12	Additional CWD factors
political_violence	15	21	**Political violence**
puig_gdp	9	13	**Emission forecasts from Mexico**
puig_oil	6	19	**Oil emissions and prices**
sanDiego	7	10	Effectiveness of surgical procedures
Sheep	14	15	Risk management policy for sheep scab control
spain	5	10	Future antimicrobial esistance in Spain
SPEED	14	16	**Volcano hazards (Vesuvius & Campi Flegrei, Italy)**
Tadini_Clermont*	12	13	Somma-Vesuvio volcanic complex geodatabase
Tadini_Quito*	8	13	**Volcanic risk**
TdC*	18	17	**Volcano hazards (Tristan da Cunha)**
Tobacco*	7	10	Grant effectiveness, childhood obesity
TOPAZ*	21	16	Tectonic hazards for radioactive wastes siting in Japan
UK	6	10	**Future antimicrobial esistance in UK**
UMD	9	11	Nitrogen removal in Chesapeake Bay
USGS	32	18	**Volcanos**
Washington	5	10	Grant effectiveness, child health insurance enrollment
* contains one variable where *PWi*MD=*true value*			
** contains 4 variables where *PWi*MD=*true value*			

Note. # Calib Vbls refers to the number of calibration variables that each expert assessed in a study. Studies are sorted in alphabetical order. Variables where *PWi*MD=*rls* are excluded in section 5, since they cannot be rendered scale invariant.

Table A3.3 References for post 2006 expert studies.

Study Name	Reference
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.
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.
ATCEP	Morales-Nápoles, O., Kurowicka, D., & Cooke, R. (2008). EEMCS final report for the causal modeling for air transport Safety (CATS) project.
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
Biol Agents	Aspinall & Associates (2006). REBA Elicitation. Commercial-in-confidence report, 26pp.
Brexit_food	Barons MJ, Aspinall W. (2020) Anticipated impacts of Brexit scenarios on UK food prices and implications for policies on poverty and health: a structured expert judgement approach. BMJ Open 2020;10:e032376. doi:10.1136/ bmjopen-2019-032376
CDC ALL	past clearance, publication in preparation
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.
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.
create	Bier V.M, Kosanoglu, F, Shin J, unpublished data, nd.
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
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.
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.
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
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
	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.
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.
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."
Geopolit	Ismail, Raveem and Ried, Scott (2015) "Ask the Experts" , The Actuary, the official magazine of the Institute and Faculty of Actuaries 6/15/2016.
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
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.
hemophilia	Fischer K, Lewandowski D, Janssen MP. Estimating unknown parameters in haemophilia using expert judgement elicitation. Haemophilia. 2013 Sep;19(5):e282-e288.
Ice 2012	Bamber, J.L., and Aspinall, W.P., (2012) An expert judgement assessment of future sea level rise from the ice sheets, Nature Climate Change, DOI: 10.1038/NCLIMATE1778. http://www.nature.com/nclimate/journal/vaop/ncurrent/full/nclimate1778.html
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, PNAS first published May 20, 2019 https://doi.org/10.1073/pnas.1817205116
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.
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
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."
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.
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.
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.
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
PoliticalViolence	In preparation
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.
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.
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).

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."
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/
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. Anew information geodatabase with uncertainty characterizations, J. Geophys. Res. Solid Earth, 122, doi:10.1002/2016JB013858.
	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
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.
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
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.
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
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."
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).
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.