Letter to Editor JRSS – A: Ganguly, T., Wilson, K.J., Quigley, J., and Cooke, R.M.

Reaction to Babuscia, A. and Cheung, K-M "An approach to perform expert elicitation for engineering design risk analysis: methodology and experimental results", J.R. Statist. Soc. A (2014)

Dear Madam/Sirs,

It is pleasing to see an article on expert judgment in JRSS, to see attention paid to validation and to see abundant references to the "Cooke classical model". The article proposes a four part methodology comprising probabilistic thinking (part 1), calibration (part 2), elicitation (part 3) and experts' aggregation (part 4) to perform expert elicitation for an application in engineering design risk analysis. In part 1, the biases and heuristics that exist within the the human brain are tackled through a questionnaire and a qualitative score is obtained. This score is based on the sum of all of the question scores: the questions are given equal weights. Part 2 uses a calibration process in which the calibration score is based on the experts' tendency to overestimate or underestimate quantities in their field of expertise rather than the experts' performance on seed variables. The calibration score is defined in Eq. (3) of this letter. In part 3 the bounds and shapes of the distributions are elicited from the experts, who are then shown the resulting distribution and declare whether it is satisfactory. If not, the quantities are re-elicited to arrive at the experts' elicited distribution. The last part of the methodology deals with mathematical aggregation and the scores from part 1 and part 2 are used to derive the aggregate distribution.

This letter (a) rectifies mis-representations of the classical model, (b) identifies methodological shortcomings in the authors' approach and (c) compares performance of the authors' "calibration weighting" with equal weights and global weights of the classical model on the large expert judgment data set made available in Cooke and Goossens (2008).

a) The classical model treats experts as statistical hypotheses and scores them with regard to statistical likelihood (known as calibration) and information. The calibration score is based on experts' assessments of "seed variables" from their field whose true values are known post hoc. Specifically, it is the p-value of falsely rejecting the hypothesis that the realizations are independently drawn from a distribution complying with the expert's assessed quantiles. Information is Shannon relative information with respect to a user chosen background measure. Shannon relative information is used because it is a familiar, scale invariant, tail insensitive slow function. The theory of proper scoring rules is invoked to compute un-normalized weights as a product of information and statistical likelihood scores. It is not the case that seed variables have "a real probability distribution", they have realizations. Describing these measures as "the scoring of each calibration question and the level of certainty that is associated with that quantity" invites confusion: calibration questions are scored collectively not individually and "level of certainty" poorly describes the information score, as it neglects the role of the background measure. Harold Jacobson (1969) (not "Harold and Jacobson (1969)") derives an upper bound on the variance of *unimodal* distributions absolutely continuous with respect to Lebesgue measure. The statement that the uniform distribution "achieves the maximum variance across the bounded distributions..." does not reflect reasons for choosing the uniform background measure in the classical model.

b) Babuscia and Cheung do not recommend one approach to obtain the weights but offer a choice of equal weights, or a combination of *quality weights* and *calibrations weights*, neither of which has a strong methodological warrant. Quality weights are derived from a test based on the experts' ability to think probabilistically. Simply because an expert's assessments comply with the laws of probability does not mean they will provide statistically accurate and informative predictions. Standard elicitation processes address biases and lack of probabilistic thinking through training, such as Spetzler et al (1975), Merkhofer (1987), Clemen and Reilly (2001) or, specifically in the context of engineering design, Walls and Quigley

(2001). The authors' calibration weights are derived from assessing the (truncated) mean squared percentage error of each expert, through comparing realizations unknown to the expert with their best estimate. A perfectly calibrated expert could be very uninformative.

Once the scores are obtained a bisection method is applied to obtain weights. This is achieved through ranking the experts and assigning weight of 0.5^{n-1} to the n^{th} ranked expert. Such an approach will always assign a weight of 0.5 to the top ranked expert regardless of the number of experts or the quality of their judgments. In a situation of two experts, each would receive equal weighting regardless of their performance. This approach to weighting is ad hoc, ignoring much of the information obtained through the elicitation exercise.

Generic performance measures defined within the classical model are statistical accuracy and informativeness. With these measures, performance based weighting out performs equal weighting (Cooke & Goossens, 2008). Other performance measures could be contemplated (Flandoli et al, 2011, Wisse et al, 2008) and weighting schemes based on these schemes also out-perform equal weights. Hora (2004) shows that equal weighting of well-calibrated experts destroys calibration. We are unaware of any performance based defense of equal weighting.

c) Data on 50 professional expert judgment studies was made available in Cooke and Goossens (2008) and has been used to evaluate several scoring proposals (see eg. Kallen and Cooke, 2002, Cooke et al, 2008, Wisse et al, 2008)). The scoring variables, like that in Eq. (3), have been studied in various student theses, with the general result that such schemes do not strongly outperform equal weighting. Such equal weighting methods do not consider any information on the performance of the experts and as such we use this as a baseline for comparison. Through their Quality Weighting scheme, Babuscia and Cheung propose a method for obtaining weights that reflect the relative differences in performance of experts through the bi-section method. The resulting weights are determined by ranked performance only ignoring much of the relative performance data available.

Two of the 50 studies were unsuitable because they had no realizations or had no medians. We have formed weighted combinations based on Eq. (3) (called BC weights) for the 48 remaining studies. These have been evaluated with respect to the statistical likelihood score (the calibration score), and also with the scoring variable in the authors' Eq. (3). In this comparison only global weights are used without adjusting the statistical power to 10 effective calibration items (for this reason the results do not always agree with the published values, but are the easiest for other researchers to reproduce)[1]. Figure 1 shows scatter plots of statistical likelihood (calibration) scores of global weights (GW) versus BC weights (BCW), BC weights versus equal weights (EW), and global weights versus equal weights. The numbers of cases for which the p-value was less than or equal to 5% were 7 for EW, 8 for BCW and 3 for GW. (NB: some data points coincide in the plots).

Babuscia and Cheung provide their own definition of "calibration" $S(e,r)$ for a vector of expert assessments e based on a vector of realizations denoted by r :

[1] See (Cooke and Goossens 2008) for details. Briefly, global weights are the same for each variable and use the experts' average information scores. Item weights or item-specific weights use the information score for each item, thereby allowing an expert to up/down weight him/herself per item. Item weights are preferred if indeed they outperform global weights, which they usually do. For cross-study comparisons, the power of the calibration scores should be equalized, and 10 effective seed items is often chosen as default.

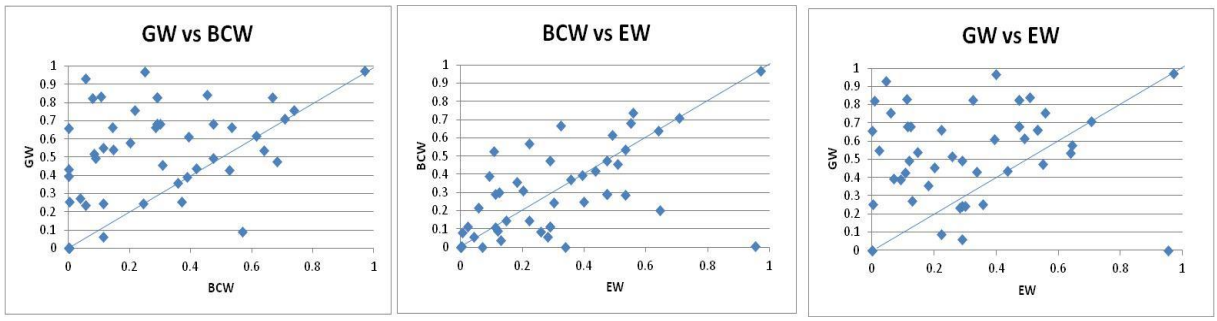

Figure 1: Scatter diagrams comparing calibration scores for the three aggregation methods showing GW strictly outperforms each the other two methods in 37 of the 48 cases.

$$S(e,r) = 100\left(1 - \frac{\sum_{i=1}^{m}\left(\frac{e_i - r_i}{r_i}\right)^2}{m}\right), \qquad |e_i - r_i| \le 2 \tag{3}$$

The expert e's best estimate for item i, e_i, is truncated at $2r_i$. Using Babuscia and Cheung's definition of "calibration" in Eq. 3, as a performance measure, BCW weighting performs better with an average over all 48 studies of 69.94 in comparison to GW (64.48) and EW (64.13). Of course, the truncation in Eq. (3) amounts to changing the experts' best estimate based on information from the realized value, which is unavailable if the realization is not known. On removing this truncation and summing the percentage error over all realizations in all studies, the sum square percentage error was lowest (i.e. best) for GW; EW and BCW were respectively 3% and 29% higher than GW.

A larger issue is raised by the decision to alter the experts' distributions based on the BC calibration information. Experts are chosen to participate in these panels because of their knowledge and standing in their fields, and the choice of experts is an important part of the message. Changing the experts' distributions blurs the distinction between expert and analyst in a way which is inimical to the classical model. Finally, basing an evaluation of a method on one variable from one study is precarious.

References

1. Clemen RT and Reilly T (2001) Making Hard Decisions with Decision Tools, Pacifc Grove, CA, Duxbury Press
2. Cooke, R.M., ElSaadany, S., Xinzheng Huang, X. (2008) On the Performance of Social Network and Likelihood Based Expert Weighting Schemes, Reliability Engineering & System Safety, 93, issue 5 745-756
3. Cooke, R.M., Goossens, L.H.J. (2008) TU Delft Expert Judgment Data Base, Special issue on expert judgment Reliability Engineering & System Safety, 93, issue 5 pp 657-674
4. Flandoli, F., E. Giorgi, W. P. Aspinall and A. Neri (2011) Comparison of a new expert elicitation model with the Classical Model, equal weights and single experts, using a cross-validation technique. Reliability Engineering and System Safety, 96, 1292-1310. doi:10.1016/j.ress.2011.05.012.
5. Hora, S.C. (2004) "Probability Judgments for Continuous Quantities: Linear Combinations and Calibration" Management Science, Volume 50 , Issue 5 (May 2004) Pages: 597 - 604 .
6. Jacobson, H. (1969) "The Maximum Variance of Restricted Unimodal Distributions", Annals of Mathematical Statistics, Vol. 40, No. 5, pp. 1746-1752
7. Kallen, M.J. and Cooke, R.M.,(2002) "Expert aggregation with dependence" Probabilistic Safety Assessment and Management E.J. Bonano, A.L. Camp, M.J. Majors, R.A. Thompson (eds),Elsevier, 1287-1294.
8. Merkhofer, M (1987) Quantifying Judgemental Uncertainty: Methodology, Experiences, and Insights *IEEE Trans. On Systems, Man. And Cybernetics,* **17,** pp741 - 752.
9. Spetzler, C and Stael von Holstein, CA (1975) Probability Encoding in Decision Analysis, *TIMS,* **22,** 340 - 358.
10. Walls L and Quigley J (2001) 'Building Prior Distributions to Support Bayesian Reliability Growth Modelling Using Expert Judgement' *Reliability Engineering and System Safety* **74** pp 117-128
11. Wisse B, Bedford T and Quigley (2008) J Expert Judgment Combination Using Moment Methods, Special issue on expert judgment Reliability Engineering & System Safety, 93, issue 5 pp 675-686