Appendix S1. SEJ methodology and out-of-sample validation.

SEJ METHODOLOGY

Introduction

Structured expert judgment involves two generic, quantitative measures of expert performance: calibration and information. A detailed description of these measures and the expert scoring procedure is available elsewhere (Cooke, 1991; Wittmann et al., 2014, 2015). Calibration measures the statistical likelihood that a set of experimental results correspond with an expert's given assessments. It is given by the p-value at which we would falsely reject the hypothesis that an expert's probability statements are accurate. Scores near zero indicate that it is unlikely that the expert's assessments are correct; higher scores indicate the expert's assessments are supported by the data. Information, the second scoring variable, is a measure of how tightly distributed an expert's assessments are relative to a background distribution. A combination of expert assessments is called a "decision maker" (DM). Calibration and information are multiplied to form an expert's combined score, which is the basis of a DM with performance weights that gives high weight to experts with good calibration and high information and zero weight to poorly calibrated experts.

Calibration

In this study the 5th, 50th and 95th percentiles were elicited from each expert for each of the continuous variables. Hence, each expert divides the range of possible outcomes of each variable into 4 intervals: less than or equal to the 5% value, greater than the 5% value and less than or equal to the 50% value, etc. The probabilities for these intervals are expressed as a vector:

$$p = (p_1, p_2, p_3, p_4) = (0.05, 0.45, 0.45, 0.05)$$

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

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

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

$$2NI(s(e) \mid p) = 2N \sum_{i=1..4} s_i \ln(s_i / p_i) \qquad (1)$$

is asymptotically distributed as a chi-square variable with 3 degrees of freedom. This is the likelihood ratio statistic, and $I(s \mid p)$ is the relative information of distribution s with respect to p.

Extracting the leading term of the logarithm yields the familiar chi-square test statistic for goodness of fit. There are advantages in using the form in equation 1 (Cooke, 1991).

If after a few realizations the expert were to see that all realization fell outside his 90% central confidence intervals, (s)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 (s)he learns from the realizations. Expert learning is not a goal of an expert judgment study. Rather, the problem owner wants experts who do not need to learn from the elicitation. Independence is not an assumption about the expert's distribution but a desideratum of the problem owner. Hence the decision maker (see below) 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 (equation 1), and the likelihood, or p-value, or **calibration score** of this hypothesis, is:

$$Cal(e) = p\text{-}value(e) = Prob\{2NI(s(e)\ p) \geq r \mid H_e\}$$

where r is the value of (equation 1) based on the observed values $x_1,\ldots x_N$. It is the probability under hypothesis H_e that a deviation at least as great as r should be observed on N realizations if H_e were true. Calibration scores are absolute and can be compared across studies. However 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 (equation 1), without changing the sample distribution s.

Although the calibration 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. High scores, near 1, indicate good support.

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" must be measured relative to some other distribution.

Measuring information requires associating a density with each assessment of each expert. To do this, we use the unique density that complies with the experts' quantiles and is minimally informative with respect to the 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. 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

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

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

$$Inf\,(e) = Average\ Relative\ information\ w.r.t.\ Background = (1/N) \sum_{i=1..N} I(f_{e,i} \mid g_i)$$

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 calibration, 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 her/himself a high information score by choosing his quantiles very close together. 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.

The above information score is chosen because it is familiar, tail insensitive, scale invariant and "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 calibration score, which is a very "fast" function. Together these features cause the product of calibration and information to be driven by the calibration score.

Combination: Decision Maker

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

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

where $1_\alpha(Cal(e) \geq \alpha) = 1$ if $Cal(e) \geq \alpha$, and is zero otherwise. The combined score thus depends on α; if $Cal(e)$ 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 true beliefs (Cooke, 1991). α is similar to a significance level in simple hypothesis testing, but its origin is to measure 'goodness' and not to reject hypotheses.

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 calibration (high

statistical likelihood, high p-value) and high information. Weights that reward good expertise and pass these virtues on to the decision maker are desired.

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 calibration and information. Solving this problem on real data, one finds that the weights do not generally reflect the performance of the individual experts. As an expert's influence on the decision maker should not appear haphazard, and "gaming" the system with assessments tilted to achieve a desired outcome should be discouraged, we must impose a strictly scoring rule constraint on the weighting scheme.

The scoring rule constraint requires the term $1_\alpha(Cal(e) \geq \alpha)$ in equation 2, but does not indicate what value of α we should choose. Therefore, we choose α to maximize the combined score of the resulting decision maker. Let $DM_\alpha(i)$ be the result of linear pooling for any item i with weights proportional to equation 2:

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

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

$$calibration\ score(DM_{a^*}) \times information\ score(DM_{\alpha^*}). \tag{4}$$

This weight is termed global as the information score is based on all the assessed seed items (i.e., calibration variables).

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:

$$w_\alpha(e,i) = 1_\alpha(calibration\ score) \times calibration\ score(e) \times I(f_{e,i} \mid g_i) \tag{5}$$

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

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

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

$$calibration\ score(IDM_a) \times information\ score(IDM_\alpha) \tag{7}$$

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

Item weights are potentially more attractive as they allow an expert to up- or down- weight her/himself for individual items according to how much (s)he feels (s)he knows about that item. "Knowing less" means choosing quantiles farther apart and lowering the information score for that item. Of course, good performance of item weights requires that experts successfully

perform this up-down weighting. 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 described as optimal weights under a strictly proper scoring rule constraint. With both global and item weights, calibration strongly dominates over information, and information serves to modulate between more or less equally well calibrated experts.

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 calibration and the information of any proposed decision maker. We should hope that the decision maker would perform better than the result of simple averaging, called the *equal weight DM* (EW), 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 weighted DM's* (PW). In general the optimized global weight DM is used, unless the optimized item weight DM is markedly superior.

General Considerations

We expect that many readers will have a healthy skepticism of the value of SEJ in estimating uncertainty. Below, we explain some key features of SEJ in the context of quantifying uncertainty in BMP performance and we clarify some common misconceptions of the approach.

1) In the present study, SEJ is a means for quantifying experts' uncertainty regarding the efficiency of stormwater BMPs; SEJ does not provide a means to estimate BMP efficiency. An estimate of BMP efficiency would be a number; the quantification of uncertainty is a distribution. Numerical estimates based on empirical measurements are accompanied with standard deviations reflecting the measurement variability, ideally based on replicated measurements under controlled conditions. If such measurements were available, there would be no need for SEJ. SEJ is designed to estimate uncertainty in a measurable quantity by evaluating expert-provided distributions against a subset of empirically measured calibration variables.

2) SEJ is not a survey. It is not intended to estimate some quantity like a mean value characterizing a population of experts. For surveys the method of sampling is critical and uncertainty in survey results reflects sampling fluctuations in the survey population. SEJ is designed to achieve an empirically validated quantification of uncertainty based on performance on calibration variables from the experts' field. The primary objective is to achieve uncertainty assessments that are statistically accurate and informative. Therefore, the main question is: "Does performance on the calibration variables predict performance on the variables of interest?" If we could address this question directly, there would be no need of SEJ in the first place. However, we can approximate the answer to this question by performing an out of sample cross validation where we assess the reliability of the resulting performance-based combination for all possible subsets of calibration variables (see below).

3) SEJ cannot produce new knowledge; only the scientific method can do that. SEJ is for quantifying the uncertainty in what is currently known and it provides a way to extrapolate current knowledge. The limitations of existing data are the reasons for conducting the SEJ in the first place; if we already had data on the degree to which some specific environmental factor affected uncertainty in BMP performance, then there would be no need to ask experts to estimate

the magnitude of that effect. In addition to published data, experts can draw on logic, first principles, and experience. Their ability to translate those factors into meaningful estimates of uncertainty is judged by their performance on the calibration variables.

4) The performance-based decision-maker weights each expert according to the statistical accuracy and informativeness of their estimates relative to a set of calibration variables. This weighting scheme may not always correspond to other metrics of 'expert quality'. For example, Cooke et al. (2008) compared citation-based weights with performance weights and found that citation weighting did not perform better than equal weighting. Similarly, expert self-weighting has been shown to be inferior to performance-based weighting (Aspinall and Cooke, 2013).

OUT-OF-SAMPLE VALIDATION

Background

Calibration and information are multiplied to form an expert's combined score, which is the basis of a DM with performance weights that gives high weight to experts with good calibration and high information and zero weight to poorly calibrated experts. It is the hope of structured expert judgment that a performance weight DM (PW DM) performs better than a simple averaging of expert assessments, or equal weight DM (EW DM). Comparing the combined score of the PW DM to that of the EW DM provides feedback on the in-sample validity of choosing the PW DM over a simple average of all experts. However, the out of sample validity of the classical model is also an important question. Because most expert elicitations are based on questions that cannot be observed directly or in a reasonable time frame, true out of sample validation of the model is typically impossible. However, cross validation methods that split the seed items (i.e., calibration variables) into training and test sets offer an alternative way to explore out of sample validity and support the choice of a DM. Here we summarize existing out of sample validation methods for expert elicitation and explore the out of sample performance of the present dataset.

Cooke and Goossens (2008) summarized a TU Delft database comprised of 45 studies in which experts assessed calibration variables. Researchers have used this resource to explore new models and to study whether good performance on the calibration variables is linked with good performance on the variables of interest. In a few studies, variables of interest were later observed, enabling true out-of-sample validation. In most cases, like the present case, the variables of interest are not observable on time scales relevant for the decision problem. Therefore, various forms of cross validation have been applied. Clemen (2008) proposed a Remove-One-At-a-Time (ROAT) method according to which the calibration variables were removed one at a time and predicted by the model initialized on the remaining calibration variables. These predictions, though originating from different DMs, were pooled and compared with the EQ DM. On 14 studies analysed by Clemen, the PW DM was superior to the equal weight DM on 9, which was not statistically significant. ROAT is biased against the PW DM, since each calibration variable was predicted by a DM in which experts who assessed that particular item badly were up-weighted, and all variables were assessed in this manner. It is commonly observed that removing one calibration variable can influence an individual expert's p-value by a factor 3 or more (Cooke, 2008), a feature explained by the fact that statistical accuracy is a very fast function. It is easy to under appreciate this effect, and Cooke (2012) gave a detailed example making this conspicuous. Variations on the ROAT approach have been performed by Lin and Cheng (2008, 2009) and Flandoli et al. (2011). The most extensive study of this kind is Eggstaff et al. (2014), which initializes the global weights model on *all* subsets of calibration variables and in each case predicts the complementary subset. Some researchers apply other scoring variables to measure performance which apply to single variables rather than sets of variables, such as the quadratic or logarithmic rule. An extensive discussion in Cooke (1991) discourages this practice.[1] Cross validation studies usually do not consider the item weight DM, which is the variant most often used in practice. In spite of these caveats, a general

[1] A simple example clarifies the issue. Suppose an expert assesses the probability of Heads as 1/2 with a coin of unknown composition. On each toss with the coin, the score is the same for Heads and Tails. If these individual scores are added, then the sum score after 100 tosses is also independent of the actual sequence of outcomes; 50 Heads and 50 Tails gets the same score as 100 Heads.

conclusion of all this work is that the performance based DM is degraded on out of sample prediction, but it is superior to the equal weight DM.

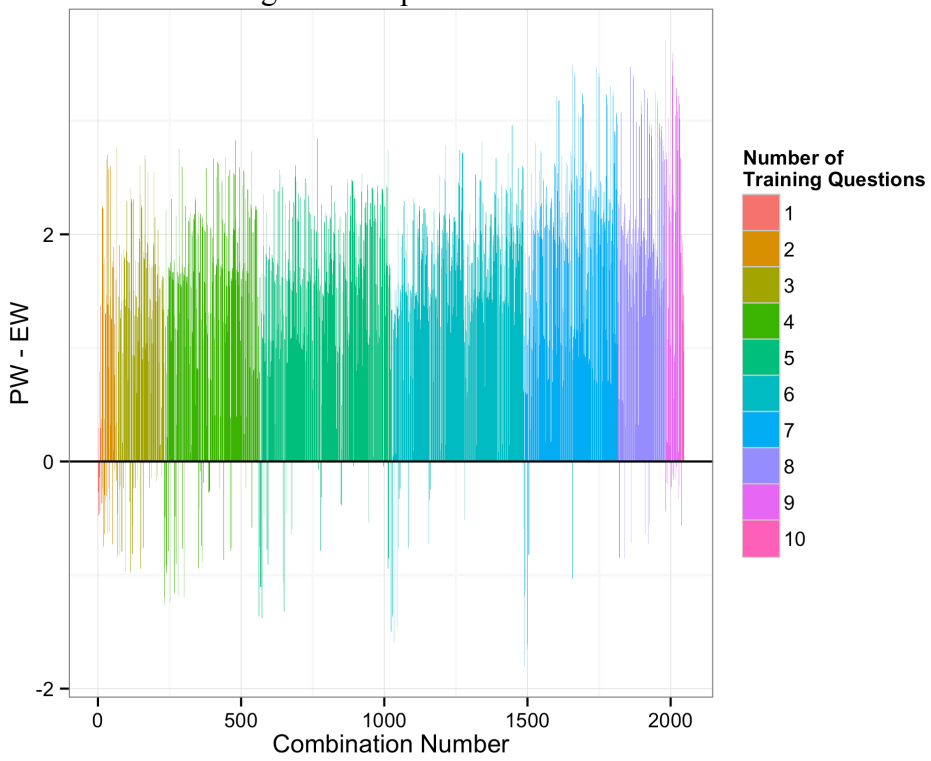
When a cross validation study initializes the PW DM on K of the N calibration variables, and measures performance on the remaining N - K variables, the following issues arise: (1) If K is close to N, then the small number of out-of-sample predictions have low statistical power and are subject to the bias noted above, and (2) If K is small, then the ability to distinguish experts' high and low statistical accuracy is low. If the experts' calibration scores on the K variables are similar, then weighting is driven by informativeness, which is often negatively correlated with statistical accuracy (Cooke, 1991). However, Eggstaff et al. (2014) noted that with small K the actual scores did not predict the scores on the larger set of calibration variables, but the superiority of performance weighting against equal weighting was attested.

Methods

For this exercise there were 10 experts and 11 seed items. All experts assessed all items. We follow the comprehensive Eggstaff method of out of sample validation. As noted previously (Cooke et al., 2014), the comprehensive approach is not optimal, but it balances the known biases of other existing methods. We thus divide the N seed items into all possible combinations of K training items and $N - K$ test items. For the 11 seeds in this study, this yields 2046 total combinations. For a given combination, expert weights are calculated based on the training set, the weights are used to predict a PW DM's assessments items in the test set, and those assessments are scored according to the established scoring procedure. EW DM assessments of the test items are also calculated and scored for comparison. As in a classical elicitation, the DMs are scored according to their calibration and information, and those two scores are multiplied to produce a combined score. Although no succinct summary measure exists to describe the out of sample performance of a given study, trends in PW DM performance versus EW DM performance across the combinations of training and test sets are analyzed for a general perspective.

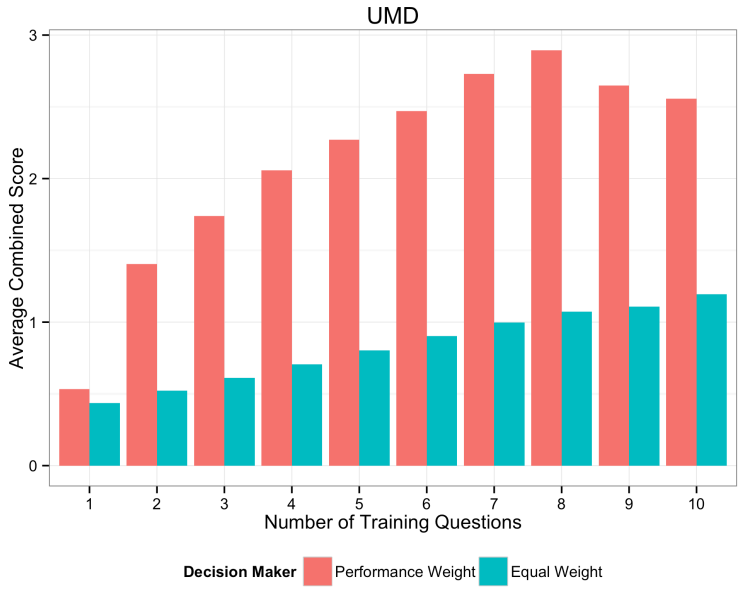
Results

Figure 1. The difference between Performance Weight and Equal Weight DM combined scores for each combination of training and test questions.



Dividing the 11 seeds from the dataset into *K* training and *N* - *K* test questions produces 2046 possible combinations. The PW DM of a given combination outperforms the EW DM (as determined by a higher combined score) 93% of the time. Over all 2046 runs, the average difference between PW and EW combined scores is 1.49. In an unpublished analysis of 31 other recent SEJ studies, no study had an average difference as high as that observed here, and in only one other study did the PW DM outperform the EW DM over 90% of the time.

Figure 2. Average Performance Weight and Equal Weight DM scores for each training set size.



Combined scores for both the PW and EW DMs increase as the size of the training set increases (Figure 2). The EW combination is the same regardless of variables in the training set or training set size, and the increase in the EW combined score is due to a loss of statistical power as the test set size shrinks. The PW DM combined score, however, increases more than that of the EW, suggesting an improvement in performance beyond the loss of statistical power when evaluating the test set. This is explained by the growing resolving power to distinguish good performing experts from poor performing experts on the test set as the test set size increases. The PW DM outperforms the EW DM for all training set sizes. Prior analysis of 62 studies found this to be true roughly two-thirds of the time (Eggstaff et al., 2014).

Figure 3. Expert combined scores for each combination of training and test questions.

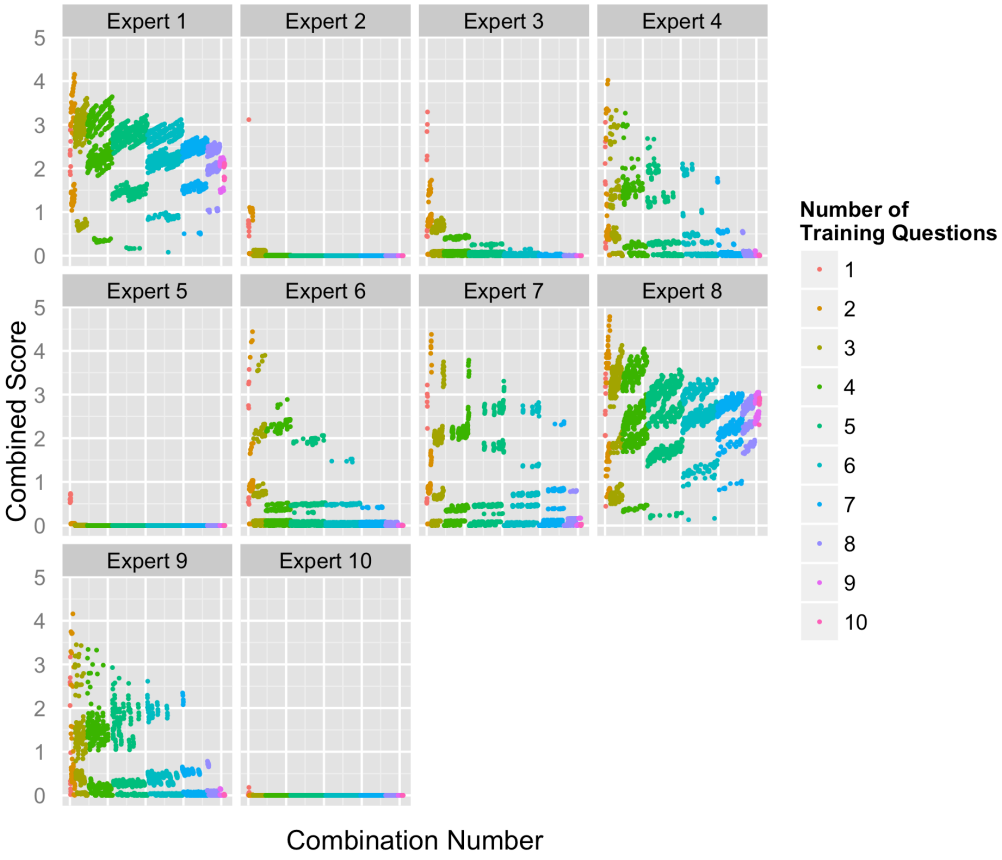
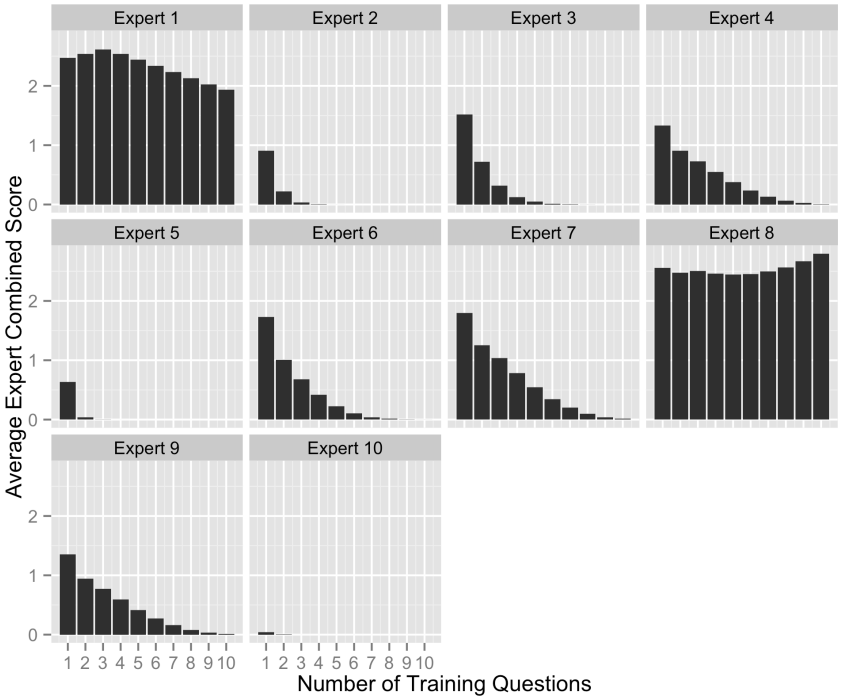
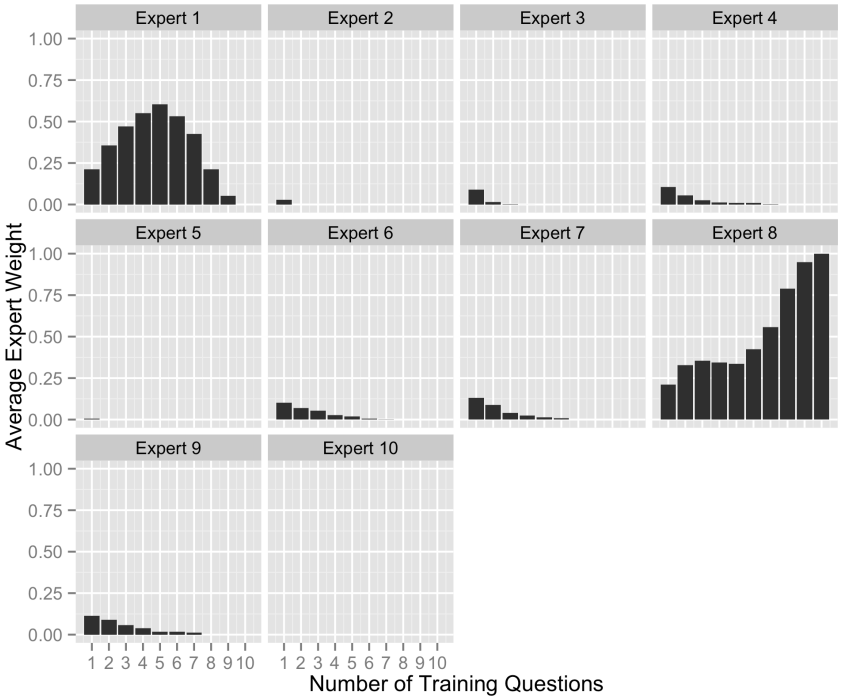


Figure 4. Average expert combined scores for each training set size.



As the training set increases, the variation in an expert's combined scores decreases (Figure 3). Average combined scores decrease toward 0 for all experts except Expert 1 and Expert 9 (Figure 4), the only experts with combined scores greater than 0.001 in the final study analysis.

Figure 5. Average expert weights for each training set size.



Although many experts receive weight in some early combinations (when there are only 1 or 2 training items used to initialize the expert scores), as the training set size increases, average combined weights approach 0 for all experts except Expert 8, the only expert weighted in the final PW DM combination in the study.

For this study, out of sample analysis demonstrates that performance-based combinations of the experts consistently outperform equally weighted combinations. This expert cohort involved a relatively large number of high scoring experts. It remains to be seen how out of sample performance changes for a case with fewer well-calibrated experts.

REFERENCES

Aspinall WP, Cooke RM. 2013. Quantifying scientific uncertainty from expert judgement elicitation, in Rougier J, Sparks S, Hill LJ eds., *Risk and Uncertainty Assessment for Natural Hazards*. Cambridge, UK: Cambridge University Press. pp. 64–99.

Clemen RT. 2008. Comment on Cooke's classical method. *Reliab Eng Syst Saf* **93**:760–765. doi:10.1016/j.ress.2008.02.003.

Cooke R. 2008. Response to discussants. *Reliab Eng Syst Saf* **93**:775–777. doi:10.1016/j.ress.2008.02.006.

Cooke RM. 1991. *Experts in uncertainty: opinion and subjective probability in science*. Oxford University Press, New York, New York.

Cooke RM. 2012. Pitfalls of ROAT cross-validation: comment on effects of overconfidence and dependence on aggregated probability judgments. *J Model Manag* **7**:20–22. doi:10.1108/17465661211208785.

Cooke RM, ElSaadany S, Huang X. 2008. On the performance of social network and likelihood-based expert weighting schemes. *Reliab Eng Syst Saf* **93**:745–756. doi:10.1016/j.ress.2007.03.017.

Cooke RM, Goossens LLHJ. 2008. TU Delft expert judgment data base. *Reliab Eng Syst Saf* **93**:657–674. doi:10.1016/j.ress.2007.03.005.

Cooke RM, Wittmann ME, Lodge DM, Rothlisberger JD, Rutherford ES, et al. 2014. Out-of-sample validation for structured expert judgment of Asian carp establishment in Lake Erie: Out-of-Sample Validation for Structured Expert Judgment. *Integr Environ Assess Manag* **10**:522–528. doi:10.1002/ieam.1559.

Eggstaff JW, Mazzuchi TA, Sarkani S. 2014. The effect of the number of seed variables on the performance of Cooke′s classical model. *Reliab Eng Syst Saf* **121**:72–82. doi:10.1016/j.ress.2013.07.015.

Flandoli F, Giorgi E, Aspinall WP, Neri A. 2011. Comparison of a new expert elicitation model with the Classical Model, equal weights and single experts, using a cross-validation technique. *Reliab Eng Syst Saf* **96**:1292–1310. doi:10.1016/j.ress.2011.05.012.

Lin S, Cheng C. 2009. The reliability of aggregated probability judgments obtained through Cooke's classical model. *J Model Manag* **4**:149–161. doi:10.1108/17465660910973961.

Lin S-W, Cheng C-H. 2008. Can Cooke's model sift out better experts and produce well-calibrated aggregated probabilities? *Proc IEEE Int Conf Ind Eng Eng Manag* 425–429. doi:10.1109/IEEM.2008.4737904.

Wittmann ME, Cooke RM, Rothlisberger JD, Lodge DM. 2014. Using structured expert judgment to assess invasive species prevention: asian carp and the Mississippi—Great Lakes hydrologic connection. *Environ Sci Technol* **48**:2150–2156. doi:10.1021/es4043098.

Wittmann ME, Cooke RM, Rothlisberger JD, Rutherford ES, Zhang H, et al. 2015. Use of structured expert judgment to forecast invasions by bighead and silver carp in Lake Erie: Forecasting Biological Invasion. *Conserv Biol* **29**:187–197. doi:10.1111/cobi.12369.