

Comment on Cooke's Classical Method

Robert T. Clemen
Fuqua School of Business
Duke University
Durham, NC 27708
USA

Email: clemen@duke.edu
Phone: +1.919.660.8005
Fax: +1.919.684.2818

November 29, 2006

Abstract

Several of the papers in this special issue are in one way or another linked to Cooke's "classical" method for combining expert probability distributions. This comment focuses on characteristics of that method. In particular, I consider two questions: Does the weighting scheme give the experts a positive incentive to report their beliefs honestly for each variable? How does Cooke's method perform when evaluated out-of-sample?

Introduction

Congratulations to all of the authors represented in this special issue of *Reliability Engineering and System Safety*. Several of the articles herein relate in one way or another to Cooke's "classical" method [1] for combining expert probability distributions, such as those that one might obtain in a risk assessment. To my knowledge, Cooke's method is the most widely used approach for combining expert judgments. The method has been applied in many different settings as can be seen in Table 2 in Cooke and Goossens' overview article in this issue

(hereafter referred to as CG). Moreover, Cooke's is the only method in use in which real data ("seed variables") are the basis for evaluating the experts and calculating weights to use in combining their probabilities.

As someone who has worked in this area, in both research and practice, I can vouch for the difficulty of developing and applying expert combination models. The topics of the papers in this issue, ranging from evaluations and demonstrations of the method to extensions and variations, demonstrate the widespread interest in Cooke's method. The other comments have provided interesting perspectives on the individual articles. Rather than commenting on the articles, I will focus my discussion on the classical method itself.

Calculating Weights and Strictly Proper Scoring Rules

Cooke's method is an example of a "linear opinion pool," a weighted arithmetic average of the experts' probability distributions. If we let $F_i(x)$ denote expert i 's probability distribution for an uncertain variable of interest (X), then the linear opinion pool $F_c(x)$ that results from combining k experts is

$$F_c(x) = \sum_{i=1}^k w_i F_i(x),$$

where the weight assigned to $F_i(x)$ is w_i , and $\sum w_i = 1$.

Although the linear opinion pool is a popular and intuitive combination method with many useful properties [1, 2], there is no method for assigning weights that is derived entirely from first principles. One can, however, interpret the weights in a variety of ways, and each interpretation lends itself to a particular way to calculate the weights [3]. Cooke's approach falls under the category of methods in which the weights are interpreted in terms of the experts' relative performance. Strictly proper scoring rules are typically used as performance measures,

because such scores give experts a positive incentive to report their probability beliefs honestly. Weighting expert judgments on the basis of relative performance on a scoring rule is perfectly reasonable, but it begs the question of what scoring rule to use, because different scoring rules lead to different weights [3]. Cooke's method calculates an expert's weight using a score that is a combination of separate calibration and information scores (see [1] and CG).

My interest is in Cooke's calibration score. This score is in the form of the p -value of a statistical test of the null hypothesis that the expert is calibrated. Using an expert's assessed probabilities for the seed variables and the realizations for these variables, one can test this hypothesis, and Cooke takes the p -value for that test as a measure of the expert's degree of calibration. Cooke carefully and correctly points out that this scoring rule is for average probabilities; that is, an expert is scored on his or her assessments for a set of uncertain variables. In addition, Cooke proves that the scoring rule is asymptotically strictly proper [1, Proposition 9.6], which CG explain as follows:

“[S]uppose an expert has given his quantile assessments for a large number of variables and subsequently learns that his judgments will be scored and combined according to the classical model. If (s)he were then given the opportunity to change the quantile values (e.g., the numbers 5%, 50%, or 95%) in order to maximize the expected weight, the expert would choose values corresponding to his/her true beliefs.”

I have no question about the correctness of this interpretation; the logic and mathematics behind it [1] are impeccable. The concern, though, is about the expert's incentive. Consider an expert who has ten assessments to make and for each variable is asked for 10th, 50th, and 90th percentiles. The expert wants to be sure that he is perfectly calibrated on these ten assessments, in the sense

that one of the realizations falls below its assessed 10th percentile, four fall between their 10th and 50th percentiles, another four fall between their 50th and 90th, and the remaining one falls above its 90th percentile. So he adopts the following strategy: For the first variable, he sets the 10th percentile extremely high, such that the realization is virtually assured to fall below the specified value. For the next four variables, he sets the 10th percentiles very low and the 50th percentiles very high, so that the four realizations are essentially certain to fall within these wide intervals. Similarly, for the next four variables he sets the 50th and 90th percentiles such that the intervals are extremely wide, and for the last variable he sets the 90th percentiles very low. (If ten variables are not enough to satisfy CG's requirement of a "large number of variables," the numbers in the example can be scaled up as much as the reader desires.)

When the realizations are reported and fall into intervals according to the expert's design, he will indeed be found to be perfectly calibrated. Moreover, before learning the realizations, if the expert were asked whether he wanted to change the probabilities associated with the empirical distribution of realizations (i.e., the average probabilities), he would undoubtedly say, "No, thank you. I am perfectly happy with the probabilities – I am satisfied that 10% of the realizations will indeed fall in the first bin, 40% in the next bin, and so on." And he would be saying that with complete honesty; changing the probabilities in any way would be inconsistent with his beliefs. This example shows that the "macro" incentive to report average probabilities honestly nevertheless allows the expert to manipulate the individual assessments in order to optimize the overall calibration score. A scoring rule for average probabilities cannot provide the "micro" incentive needed for an expert to report honestly on each variable.

One objection to this argument is that such an expert would score very low on the information portion of Cooke's score, and thus his weight would be reduced. This argument is

correct; a tradeoff clearly exists between the calibration score and the information score, such that the expert can improve calibration through a strategy like the one described above at the expense of the information score. Presumably there is an optimum tradeoff, whereby the expert would specify the individual assessments to maximize the overall score. But is it the case that the expert's incentive to reduce entropy by reporting appropriately narrow intervals exactly offsets the incentive to manipulate the calibration score by broadening the intervals as described above, such that the net effect is that the expert should report his probability distribution honestly for each seed variable? To say it formally, let r_i denote the $(m \times 1)$ vector of quantiles reported by the expert for the i th seed variable, and let q_i denote the expert's true beliefs regarding these quantiles. Given N seed variables, $R = (r_1, \dots, r_N)$ and $Q = (q_1, \dots, q_N)$ are $(m \times N)$ matrices that represent the expert's quantile reports and true beliefs, respectively, for the N seed variables. Writing the expert's overall score as $S(\text{calibration score}, \text{information score})$, the question is whether

$$\arg \max_R [S(\text{calibration score}, \text{information score})] = Q.$$

In principle, this question could be answered, but it may prove to be intractable.

My argument above could be construed as an argument in favor of using scoring rules for individual variables. Cooke [1] argues that such scoring rules, originally designed as probability elicitation mechanisms, cannot be interpreted in any meaningful way – and hence are inappropriate for calculating expert weights – unless one knows the number of assessments made and the overall score's sampling distribution. Cooke's approach using average scoring rules does allow comparisons in situations where scoring rules for individual variables could not be used. Regardless, strictly proper scoring rules for individual variables can be interpreted in terms of average probabilities [4] and provide a positive incentive for the expert to report honestly for

each variable. Moreover, such scoring rules can yield meaningful weights [3], provided the experts have assessed probabilities for the same seed variables, as is the case for many of the studies listed in CG.

Does my concern about the propriety of Cooke's method have serious implications for practice? Perhaps in some instances, but my own experience suggests that experts who become engaged in the assessment process want nothing more than to express their beliefs as clearly and accurately as they can. Nevertheless, given the widespread interest in Cooke's method, it seems appropriate for risk analysts to understand as much as possible about the method's properties.

Out-of-Sample Performance

CG's Table 2 lists 45 risk-assessment studies that have been performed using Cooke's method. Both the number of studies and the diversity of topics are impressive. The performance measures show that in most cases the classical method ("performance weights", hereafter PW) performs better than either equal weights (EQ) or the best expert (BE). In one case (#13) EQ performs best, and in 2 cases (#10, #22) BE performs best. In 15 cases, PW and BE have the same scores; these are cases in which all experts but one are so miscalibrated that their calibration scores fall below a critical level, and hence they receive zero weight. The remaining expert receives weight 1, making PW the same as BE.

All of the performance scores reported in CG's Table 2 are calculated within-sample. That is, weights are calculated on the basis of the available data, and then performance scores are calculated using the same data. Moreover, as explained in CG and [1], the classical method includes a step in which the critical calibration threshold (α) is chosen to maximize PW's overall

score. Thus, it should come as no surprise that PW performs so well relative to the other two combination methods when evaluated within-sample.

Common practice in the analysis of forecasting methods is to evaluate and compare performance of methods using out-of-sample data. Doing so provides a more authentic picture of a method's performance than does within-sample evaluation. In risk assessment especially, decision makers should care about a method's performance on the seed variables only to the extent that it accurately reflects performance on the variables of interest, for which realizations are not known.

If we can perform an out-of-sample analysis of Cooke's method, what would we hope to learn? The crucial question is the relative performance of EQ and PW evaluated out-of-sample (PW*). If the within-sample results in CG hold for out-of-sample evaluation, then we would expect to find that PW* is better than EQ. However, much empirical literature in forecasting and risk assessment [2, 5, 6, 7, 8] suggest that EQ will perform well relative to PW*. A related question is whether some minimal number of seed variables is required to have (statistical) assurance that PW* performs better than EQ. From CG's Table 2, we see that the effective number of seed variables in the studies ranges from five (#38, MVO seeds) to 47 (#30, Dike ring failure). Of the 45 studies, 23 have ten seeds or fewer, and all but nine have less than twenty. Ten is a small number of seeds, and an important question is whether it is large enough to permit satisfactory evaluation of the experts and calculation of the combining weights.

Even with a small number of seed variables, it is still possible to get an idea of out-of-sample performance by using a resampling method [9]. In this case, we can use a simple "leave-one-out" procedure: for a given dataset with N seed variables,

Start: Set $i = 1$

1. Exclude seed variable i .
2. Calculate performance weights based on the remaining $N - 1$ seed variables.
3. Record the combined distribution (quantiles) for seed variable i using weights calculated in step 2.
4. Set $i = i + 1$.
5. If $i \leq N$, return to step 1 and repeat.
6. After collecting the N combined distributions, calculate the score for this set of distributions.

The N combined distributions constitute PW^* ; each combined distribution i is calculated using weights derived from the other $N - 1$ seed variables, thereby producing an out-of-sample combined forecast for seed variable i . A similar procedure can be used to estimate the out-of-sample performance of BE. On each iteration, identify the expert that performs best on the $N - 1$ seed variables, and choose that expert's quantiles for seed variable i as the out-of-sample best expert's (BE^*) probability distribution for that variable. The scores for PW^* and BE^* provide an indication of their out-of-sample performance. EQ uses no data, so CG's reported scores for EQ are directly comparable to the scores for PW^* and BE^* .

Cooke graciously provided the data for 36 of the studies listed in CG's Table 2, along with guidance in performing the analysis using the Excalibur software (version 1.0 light for Windows, also provided by Cooke), thereby ensuring that my results are comparable to CG's. Although various data considerations as well as time constraints precluded analyzing all 36 data sets, Table 1 shows results for 14 of them. These 14 are not a random sample; I chose these data sets in order to obtain a range of topics and sample sizes. With that caveat, it will nevertheless prove instructive to calculate some statistics on the basis of this convenience sample.

Table 1 shows calibration, information, and combination scores for EQ, PW*, and BE*. The scores for EQ are also presented in CG's Table 2, and the reader can verify the successful replication of CG's reported scores. Separating Table 1 into calibration, information, and combination facilitates comparisons; for example, comparing PW* and EQ, PW* has worse median calibration and better median information scores. PW*'s combination score is slightly better than EQ's, but the two are not significantly different (Wilcoxon $p = 0.48$), suggesting that PW* has no statistical advantage over EQ. In contrast, BE* has a reasonable information score but fails utterly on calibration and hence on the combination score.

On another dimension, the interquartile range for the combination score is much larger for PW* than for EQ. Figure 1 compares PW* and EQ combination scores, and the difference in variability is easily seen. One possible explanation for this pattern of results is as follows: PW* tends to put the bulk of the weight on one or two experts. If those experts perform as well out-of-sample as within-sample, then PW*'s score can be quite good. However, if PW* identifies experts whose good performance within-sample does not extend out-of-sample, then PW*'s score can be very low. In contrast, EQ averages all of the expert distributions, and in so doing tends to average out any extreme distributions, thereby leading to less variability in EQ's scores.

Figure 1 orders the studies by increasing number of seed variables, which allows us to consider whether there may be some threshold number of seed variables beyond which PW* performs consistently better than EQ. Unfortunately, in this small sample of 14 studies, no conclusive trend in the relationship between PW* and EQ is apparent. In the seven smaller studies (12 seeds or less), PW* is better than EQ four times. In the seven larger studies (14 seeds or more), PW* is better than EQ five times.

We can also ask whether PW* tends to be more accurate than EQ, in the sense that PW*'s median is closer to the realization, as measured by the absolute difference. Because the seed variables tend to have different units and sometimes vastly different scales, it would be meaningless to calculate an accuracy measure like the mean absolute deviation for each method and then compare them. However, we can devise a simple, scale-free method. In a given study, we can count the number of items for which each method has the lower absolute difference and use these counts to test the null hypothesis that PW* is at least as accurate as EQ. In the spirit of Cooke's calibration score, we will take the p -value from a one-tailed sign test of the null hypothesis as a measure of the relative accuracy of PW* compared to EQ. Because $p \geq 0.5$ corresponds to PW* being more accurate on 50% or more of the items, we can say that values above (below) 0.5 indicate that PW* (EQ) is more accurate. Table 2 demonstrates the calculations for the PM25 study. The p -value of 0.39 indicates that EQ is slightly more accurate than PW* in this particular study.

Table 3 shows relative accuracy measures for each of the 14 studies. In only four of the studies is PW* is more accurate than EQ. Of the remaining ten, three are ties, and four indicate that EQ is "significantly" more accurate than PW*, in the sense that the p -values are less than 0.10. These results suggest that, overall, PW* is less accurate than EQ.

Conclusion

My arguments above can be summarized as saying that Cooke's method may not have all of the qualities that we would like it to have. On one hand, the scoring method's incentive properties regarding average probabilities do not extend to incentives to report probabilities honestly for individual variables. On the other hand, the overall out-of-sample performance of

Cooke's method appears to be no better than that of equal weights; the two methods have similar median combination scores, but equal weights has less variability and better accuracy. We have not provided evidence for a threshold number of seed variables that would ensure the performance of Cooke's method. However, if such a threshold does exist, it seems safe to say that it is probably greater than 10, which from CG appears to be the modal number of seed variables used.

My out-of-sample analysis covers fewer than 1/3 of the available data sets. Given the results above, analyzing all 45 data sets is called for. In addition, Cooke's performance score and my proposed ad-hoc location measure are only two of many possible metrics that could be used to evaluate the methods. Other metrics include alternative measures of accuracy, calibration, and bias, as well as conventional scoring rules and decompositions thereof [10, 11, 12, 13, 14, 15].

Although I have raised some issues with Cooke's method, the only operational alternative I have to offer in its place is equal weights. The results here as well as elsewhere [2, 5] suggest that equal weights – a simple, arithmetic average of the assessed probability distributions – is a reasonable combining method that performs well empirically compared to other methods when evaluated in genuine out-of-sample forecasting or assessment tasks. Having spent much of my career studying various combination methods, it has been somewhat frustrating to consistently find that the simple average performs so well empirically. I have felt somewhat like the stock investor whose carefully chosen portfolio beats the overall market in some years but not in others; over the long haul a market index fund would have performed just as well. That said, equal weights has a number of advantages. It is readily understood, no data are needed to calculate weights, software needs are minimal, and all experts have an equal voice in the combined distribution.

Why equal weights performs well empirically is not hard to understand. When used to combine probability distributions for continuous variables, the method has two positive effects. First, averaging tends to cancel out location errors; if one expert's distribution is too low relative to the actual and another expert's is too high, the equal weights combination will tend to have a median that is closer to the actual than either of the individual medians. Second, it is well known that individuals, including experts, tend to be overconfident, resulting in probability distributions that are too narrow [16]. The results reported by Lin and Bier in this issue confirm this result; although there are exceptions, the majority of the experts in the risk-assessment studies considered here are overconfident. When multiple overconfident distributions are averaged, the combined distribution tends to be more spread out, and hence less overconfident, than any of the individual distributions. In fact, if one were to average distributions from a set of calibrated experts, the resulting combined distribution would tend to be underconfident [17].

If the papers in this issue are any indication, Cooke's method is widely used and of interest to many analysts as a way to combine probability distributions in risk assessment. I hope that my comments will help users understand the method a little better. Furthermore, I hope this work will lead to additional empirical research to confirm or disconfirm the results presented above.

Acknowledgements

I am grateful to Roger Cooke and Robert Winkler for many stimulating discussions on the topics presented here, and especially to Cooke for providing his data, software, and guidance in using them. This work was supported by the National Science Foundation under Grant No. SES-

0317867. Any opinions, findings, conclusions, or recommendations expressed herein are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

- [1] Cooke RM. Experts in uncertainty: Opinions and subjective probability in science. New York: Oxford University Press; 1991.
- [2] Clemen RT, Winkler RL. Combining probability distributions from experts in risk analysis. *Risk Analysis* 1999;19:187-203.
- [3] Genest C, McConway KJ. Allocating the weights in the linear opinion pool. *Journal of Forecasting* 1990;9:53-73.
- [4] Schervish MJ. A general method for comparing probability assessors. *Annals of Statistics* 1989;17:1856-1879.
- [5] Clemen RT. Combining forecasts: A review and annotated bibliography (with discussion). *International Journal of Forecasting* 1989;5:559-583.
- [6] Makridakis S, Andersen A, Carbone R, Fildes R, Hibon M, Lewandowski R, Newton J, Parzen E, Winkler R. The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting* 1982;1:111-153.
- [7] Makridakis S, Chatfield C, Hibon M, Lawrence M, Mills T, Ord K, Simmons L. The M-2 competition: A real-time judgmentally based forecasting study. *International Journal of Forecasting* 1993;9:5-22.

- [8] Makridakis S, Hibon M. The M3-competition. *International Journal of Forecasting* 2000;16,:451-476.
- [9] Good P. *Resampling methods: A practical guide to data analysis*. Boston: Birkhauser, 2005.
- [10] DeGroot MH, Fienberg SE. Assessing probability assessors: Calibration and refinement. In: Gupta SS, Berger JO, editors, *Statistical decision theory and related topics III*, vol 1. New York: Academic; 1982. p. 291-314.
- [11] Murphy A. Scalar and vector partitions of the probability score: Part I. Two-state situation. *Journal of Applied Meteorology* 1972;11:273-282.
- [12] Murphy A. Scalar and vector partitions of the probability score: Part II. *N*-state situation. *Journal of Applied Meteorology* 1972;11:1183-1192.
- [13] Matheson JE, Winkler RL. Scoring rules for continuous probability distributions. *Management Science* 1976;22:1087-1096.
- [14] Yates JF. Analyzing the accuracy of probability judgments for multiple events: An extension of the covariance decomposition. *Organizational Behavior and Human Decision Processes* 1988;41:281- 299.
- [15] Yates JF. Subjective probability accuracy analysis. In: Wright G, Ayton P, editors, *Subjective probability*. Chichester, England: Wiley; 1994. p. 381-410.
- [16] Lichtenstein S, Fischhoff B, Phillips D. (1982). Calibration of probabilities: The state of the art to 1980. In: Kahneman D, Slovic P, Tversky A, editors, *Judgment under uncertainty: Heuristics and biases* Cambridge: Cambridge University Press; 1982. p. 06-334.
- [17] Hora SC. Probability judgments for continuous quantities: Linear combinations and calibration. *Management Science* 2004;50:597–604.

Table 1. Calibration, Information, and Combination Scores for EQ, PW*, and BE* over 14 Risk Assessment Studies.

Data set		CALIBRATION			INFORMATION			COMBINATION		
ID#	Name	EQ	PW*	BE*	EQ	PW*	BE*	EQ	PW*	BE*
2	Crane Risk	0.500	0.210	0.000	0.690	1.468	0.640	0.345	0.308	0.000
4	Space Debris	0.900	0.780	0.000	0.158	0.319	1.520	0.142	0.249	0.000
5	Composite Materials	0.120	0.520	0.000	0.929	1.646	1.868	0.111	0.856	0.000
7	Risk Management	0.324	0.827	0.827	0.745	1.231	0.220	0.241	1.018	0.182
9	Dispersion Panel TUD	0.710	0.160	0.360	0.715	0.989	1.532	0.508	0.158	0.552
12	Acrylonitrile	0.280	0.001	0.006	1.511	3.516	0.911	0.423	0.004	0.005
14	Sulfur Trioxide	0.240	0.310	0.010	2.330	4.365	0.141	0.559	1.353	0.001
15	Water Pollution	0.350	0.100	0.160	1.468	1.986	2.060	0.514	0.199	0.330
16	Dispersion Panel	0.150	0.130	0.001	0.862	1.119	1.136	0.129	0.145	0.001
17	Dry Deposition	0.001	0.520	0.520	1.184	1.339	1.339	0.001	0.695	0.696
19	Wet Deposition	0.001	0.140	0.001	0.726	0.458	0.520	0.001	0.064	0.001
25	Moveable Barriers	0.220	0.005	0.001	0.570	1.311	0.792	0.125	0.007	0.000
33	Operations Risk	0.338	0.147	0.147	0.322	0.840	0.903	0.109	0.124	0.133
35	PM25	0.645	0.514	0.028	0.542	0.897	1.063	0.350	0.461	0.030
Median		0.302	0.185	0.008	0.736	1.271	0.987	0.192	0.224	0.003
Interquartile range		0.295	0.386	0.156	0.521	0.682	0.797	0.290	0.507	0.169

Table 2. Example Calculation of Accuracy Statistic for PM25 Study.

Item	Medians		Realization	Accuracy Indicator
	PW*	EQ		
1	35.17	18.54	10	-1
2	190.10	216.50	298	-1
3	20	19.35	22	1
4	168.2	202.1	241	-1
5	219.6	193.4	239	1
6	50	39.26	27	-1
7	1.015	1.013	0.9066	-1
8	1.022	1.02	0.957	-1
9	1.024	1.03	1.01	1
10	1.021	1.043	1.071	-1
11	0.9775	0.974	1.105	1
12	0.9663	0.9602	1.124	1
			Positives	5
			Negatives	7
			Sign test p -values (1-tail)	0.39

Table 3. Accuracy Measures for 14 Risk Assessment Studies.

	Accuracy Measure (PW* relative to EQ)
Crane Risk	0.39
Space Debris	0.28
Composite Materials	0.02
Risk Management	0.50
Dispersion Panel (TUD)	0.57
Acrylonitrile	0.01
Sulfur Trioxide	0.66
Water Pollution	0.50
Dispersion Panel (EUNRC)	0.50
Dry Deposition (EUNRC)	0.09
Wet Deposition	0.82
Moveable Flood Barriers	0.79
Bank Operations Risk	0.00
Particulates (PM25)	0.39
Median	0.44

Figure 1. Combination Scores for PW* and EQ and Number of Seed Variables for 14 Risk Assessment Studies.

