Roger Cooke: Response to discussants

Reply to Simon French:

This is a good moment to acknowledge many fruitful discussions with Simon French during the formative period of the classical model. The theory of asymptotically proper scoring rules for average probabilities had been worked out, but we didn't really know what to do with the significance level. This theory told us that there should be a positive cut-off level, but of course did not say what that level should be. Simon French suggested choosing this level by optimizing the combined score of the DM. Some practitioners, most notably W. Aspinall, prefer to set this level low enough that all experts receive positive weight. This is not really consistent with the scoring rule philosophy, but has worked well in practice. To ensure comparability with other results, optimization has been applied to Aspinall's data in the results reported here.

With regard to the social network weights, French’s suggestions are worthwhile and indicate that implementation is not as straightforward as might appear at first sight. I would caution that using *all* citations, just citations from the group of experts, would involve much more work.

Reply to Bob Clemen:

I am very grateful for Bob Clemen's many thoughtful remarks and for the considerable effort that has gone into producing his results. This is an excellent example of how this data can be used to advance the discussion of various weighting schemes. I break this response down into four topics:

Scoring rules

Theoretically, I have nothing to add to the discussion in (Cooke 1991). The goals of rewarding individual elicitations and obtaining statistically accurate and informative combinations are quite different. To illustrate just how different, the following table shows the quadratic score (positive sensed on [-1, 1]) for two hypothetical experts giving 1000 next day probabilities of rain. The forecasts are thrown into probability bins ranging from 5% to 95%. The scores for each prediction are summed and divided by the total number of predictions.

Table 1: Two experts assessing next day probability of rain on 1000 days

Probability of Rain next day:		5%	15%	25%	35%	45%	55%	65%	75%	85%	95%	Totals
expert 1	assessed	100	100	100	100	100	100	100	100	100	100	1000
	realized	5	15	25	35	45	55	65	75	85	95	500
expert 2	assessed	100	100	100	100	100	100	100	100	100	100	1000
	realized	0	0	0	0	0	100	100	100	100	100	500

Quadratic score expert 1 = 0.665; Quadratic score expert 2 = 0.835

Both experts are equally informative in the sense that they both attribute 5 % probability to one hundred next days, etc. Expert 1 is statistically perfectly accurate, whereas expert 2 is wildly inaccurate statistically. Nevertheless, expert 2 receives a better score.

Inter-quartile ranges

There are many ways to measure informativeness. The Shannon relative information enjoys advantages mentioned in the paper, which the inter-quartile range, or 90% confidence interval does not. But of course, for a specific purpose some other measure may be more suitable.

Median

The goal is to find statistically accurate and informative distributions, not point estimates. Good Bayesians know that point estimates require loss functions. Nonetheless, some years ago we experimented on scale invariant scoring rules based on the proximity of the median to the realization, in the spirit of Clemens suggestion. A DM based on such a scoring rule did have statistically better medians, but did not do so well with regard to statistical accuracy or informativeness. There was no significant difference between the equal weight and performance based DM's with regard to median-realization proximity (Rosland, 1996). The propriety of median-based scoring rules is of course an issue.

Out of Sample

This is the most important issue raised by Clemen: does the performance of the performance-weighted DM (PWDM) persist beyond the set of seed variables. Clemen believes that there is no significant difference between the performance weight DM (PWDM) and the equal weight DM (EWDM) outside the variables on which PWDM has been constructed.

As noted above PWDM does use optimization to remove a degree of freedom in the definition of the classical model. On every study we routinely perform robustness analysis by removing seed variables (and experts) one at a time and re-computing PWDM It is not uncommon to see the calibration scores of PWDM fluctuate by a factor 2 or 3 on ten seed variables (compare table 4 of Cooke and Goossens, this volume).

Out-of-sample validation involves basing PWDM on an initial set of seed variables, then using this PWDM on other variables and comparing performance with EWDM on these other variables. This corresponds to the way PWDM is actually used. We can do this by splitting the set of seed variables into two halves, initializing the model on one half and comparing performance on the other half. Of course, this requires a relatively large number of seed variables. There are 14 studies with at least 16 seed variables. One of these, "TNO dispersion", eluded conversion to the format of the windows software and cannot currently be read. That leaves 13 studies. Dividing the seed variables in half gives two validation runs, using the first half to predict the second and conversely. Note that the variables on which the PWDM is initialized in these two runs are disjoint. The item weight PWDM could not be computed without writing new code, so the choice of item versus global weights is denied the PWDM on this exercise.

The data from the 13 studies are shown in Table 2 below. In 20 of the 26 studies the out-of-sample PWDM out-performs EWDM. The probability of seeing 20 or more “successes” on 26 trials if PWDM were no better than EWDM is 0.0012..

Table[1] 2: 26 out-of-sample validation runs; best performer is in bold. E1 denotes the EWDM on the first half of the seed variables, E2 denotes EWDM on the second half. PW(2)1 denotes the PWDM constructed on the second half, predicting the first half, and PW(1)2 denotes the PWDM constructed on the first half predicting the second half.

Study	DM	calibration	information	Combination
TUD disper	**e1**	**0.42**	**0.646**	**0.2713**
	PW(2)1	0.21	0.8744	0.1836
	e2	**0.39**	**0.7844**	**0.3059**
	PW(1)2	0.005	1.525	0.007624
TUD depos	e1	0.52	1.119	0.5819
	PW(2)1	**0.52**	**1.42**	**0.7382**
	e2	**0.73**	**1.324**	**0.9669**
	PW(1)2	0.59	1.374	0.8108
Operrisk	e1	0.429	0.2793	0.1198
	PW(2)1	**0.5337**	**0.5749**	**0.3068**
	e2	0.5337	0.3646	0.1946
	PW(1)2	**0.185**	**1.109**	**0.2053**
dikering	e1	0.025	0.7386	0.01846
	PW(2)1	**0.4**	**0.3859**	**0.1544**
	e2	0.025	0.7814	0.01954
	PW(1)2	**0.05**	**0.6451**	**0.03225**
Thermbld	e1	0.07	0.1424	0.009967
	PW(2)1	**0.48**	**0.5527**	**0.2653**
	e2	0.005	0.1424	0.0007119
	PW(1)2	**0.07**	**0.7305**	**0.05113**
realest	e1	0.05	0.179	0.008948
	PW(2)1	**0.33**	**0.8572**	**0.2829**
	e2	0.18	0.1676	0.030168
	PW(1)2	**0.35**	**0.6724**	**0.2353**
EuDis	e1	0.52	0.9662	0.5024
	PW(2)1	**0.52**	**1.232**	**0.6408**
	e2	0.02	0.749	0.01498
	PW(1)2	**0.08**	**1.204**	**0.09635**
PIntDos 6exp. 39 items	e1	0.001	1.108	0.0011089
	PW(2)1	**0.11**	**1.038**	**0.1141**
	e2	0.23	0.3262	0.07502
	PW(1)2	**0.44**	**0.6748**	**0.2969**

[1] PintDos involved 55 seed items, and 8 experts, but two experts assessed only a small number of seed variables. The other experts’ seed assessments did not wholely overlap; 6 experts assessed 39 common seed variables used for this exercise. Similarly, AOT was restricted to 6 experts who assessed 20 common items. The Gas study was split into a corrosion and an environment panel. Many environment experts were also corrosion experts and their corrosion seed assessments were used in the original study. For this exercise only the environment seeds were used for the environment panel. In the Dikering study the multiple measurements from each measuring station were split.

	e1	0.001	0.3638	0.0003638
soil	**PW(2)1**	**0.001**	**0.4135**	**0.0004135**
	e2	0.0001	1.539	0.0001539
	PW(1)2	**0.0001**	**1.551**	**0.0001559**
	e1	0.0001	1.235	0.0001235
Gas	**PW(2)1**	**0.06**	**2.01**	**0.1206**
Environ	e2	0.72	1.274	0.9171
	PW(1)2	**0.73**	**2.342**	**1.71**
	e1	0.1	0.2046	0.02046
AOT	**PW(2)1**	**0.1**	**0.6685**	**0.06685**
6 exp	e2	0.5	0.1793	0.08964
20 items	**PW(1)2**	**0.7**	**0.5799**	**0.4059**
	e1	**0.11**	**0.6611**	**0.07272**
EU	PW(2)1	0.0001	2.048	0.0002048
WD	**e2**	**0.04**	**0.7983**	**0.03193**
	PW(1)2	0.04	0.7743	0.03097
	e1	**0.75**	**0.2427**	**0.182**
estec-2	PW(2)1	0.43	0.3623	0.1558
	e2	0.68	0.07269	0.04943
	PW(1)2	**0.35**	**0.1893**	**0.06627**

Clemen reports results on 14 validation studies that are somewhat more pessimistic (9 “success” on 14 trials; I checked half of them and verified his numbers). His method involves removing seed variables singly, computing PWDM on the remaining seeds, and using this PWDM to predict the eliminated seed. On a study with 10 seed variables there are thus 10 *different* PWDM’s. Each pair of the 10 DM’s share 8 common seeds. The criteria for selecting the 14 studies are not specified. It is difficult to see how all these factors would affect the results. Perhaps the following reasoning partially explains Clemen's less optimistic result: With a small number of seeds, removing one seed favors experts who assessed *that* seed badly and hurts experts who assessed *that* seed well, thus tilting PW* toward a bad assessment of *that* seed. This happens on *every* seed thus cumulating the adverse effect on PW*. This does not happen when *one* PWDM predicts the entire out-of-sample set of seeds. In any case, Clemen's method is not the same as picking *one* PWDM and comparing it on new observations with the EWDM.

It is not obvious that weights derived from a given scoring rule should perform best with respect to that scoring rule. Consider Figure 5 of (Cooke, ElSaadany and Huang, this volume). There we see that the DM formed from likelihood weights does not have better likelihood scores than PWDM and EWDM.

Clemen closes with a fetching metaphor of the index funds (EWDM) which the stock market guru’s (PWDM) never seem able to beat. I fetch a different metaphor. An expert viniculturalist (PWDM) can mix different grapes to produce excellent wine, but would you expect to enhance performance by mixing bottles of wine (EWDM)?

References

Cooke, (1991) Experts in Uncertainty, Oxford University Press.

Rosland, G. (1996) “Expert judgment and performance-testing of decision makers” Department of Statistics, Stochastics and Operational Research, TU Delft, May, 1996.