

Response to Comments on Lin and Bier (2007)

We thank Tony O'Hagan and Simon French for their useful comments on our article, "A Study of Expert Overconfidence." We have some specific responses to their individual comments.

Response to Comments by Tony O'Hagan

O'Hagan focuses mainly on the validity of our analysis regarding the random effects for questions and realizations, as well as the interpretations of those effects. In particular, he has pointed out that in some cases, question wording rather than question difficulty could be contributing to a high or low calibration score for particular questions. We agree that the difficulty effect in the data set we used may have been confounded with question wording, and that not all questions may have been equally clear and explicit. Therefore, the calibration scores of the questions may not reflect only question difficulty. However, there is no evidence that this contributed to the observed calibration results; also, people may differ as to what constitutes unclear wording. In any case, it would be hard to address this issue without new laboratory experiments, in which one could separately vary question clarity and difficulty.

The second issue raised by O'Hagan relates to the dependence between the judgments of different experts—in other words, the possibility of correlation among their forecasts for a given question. This could in principle account for some of the observed realization effects. For example, if all experts give similar interval estimates for some particular question, then the percentage of intervals into which any given realization falls will tend to be close to 0% or 100%, potentially challenging our interpretation of the realization effects we observed.

To explore this question, we performed a diagnostic check to confirm the normality of our observed effects. If the percentages of intervals into which the realizations fall tend to be close to 0% or 100%, as hypothesized by O'Hagan, then the normality assumption in our model will not hold, and the parameters obtained from our model fitting will not be reliable. In particular, we used normal probability plots and statistical tests to check the validity of the normality assumption in our model.

When we used a normal plot (as well as the Shapiro-Wilk and Kolmogorov-Smirnov tests) to check the normality of our estimated expert random effects, we found no

significant violations. The normal plots of the question and realization random effects did indicate some modest possible violations from the assumption of normality. In particular, we found heavier tails than expected in the normal plots. The estimates contributing to these heavy tails seemed to come primarily from the Dike Ring study. However, with these few exceptions, there did not seem to be a problem with non-normality in general.

O'Hagan's basic point (that dependence between the judgments of different experts can lead to an exaggeration of the random effects) is in principle valid. The results of our diagnostic checks indicate that any such dependence is not sufficiently great to invalidate the general conclusions of our analysis, since the data seem to agree reasonably well with the assumptions of our model. However, it would be highly desirable to address this issue more fully through collection of additional data (in particular, larger numbers of questions with multiple realizations).

Response to Comments by Simon French

We appreciate French's favorable comments about our work. One challenge raised by French is that the public may lose trust in the use of expert opinion if they see results similar to those presented in Figures 1 through 4 of our article, showing rather poor calibration in some cases. Interestingly, Bolger and Önkal-Atay (2004) suggest that "a higher goal than accurately capturing the variance of series by means of one's confidence intervals is a communicative one: to inform others of the uncertainty in the series, and thus in the forecast itself." Thus, in some cases, experts may provide useful bounds or ranges, even if not well calibrated. However, we agree with French that further study of how both decision makers and members of the general public perceive the use of expert opinion in practice would be worthwhile.

Reference

Bolger, Fergus, and Dilek Önkal-Atay (2004), "The effects of feedback on judgmental interval predictions," *International Journal of Forecasting*, 20, 29 – 39