

Multiple Experts vs. Multiple Methods: Combining Correlation Assessments

Robert L. Winkler, Robert T. Clemen

Fuqua School of Business, Duke University, Durham, North Carolina 27708-0120
{rwinkler@mail.duke.edu, clemen@mail.duke.edu}

Averaging forecasts from several experts has been shown to lead to improved forecasting accuracy and reduced risk of bad forecasts. Similarly, it is accepted knowledge in decision analysis that an expert can benefit from using more than one assessment method to look at a situation from different viewpoints. In this paper, we investigate gains in accuracy in assessing correlations by averaging different assessments from a single expert and/or from multiple experts. Adding experts and adding methods can both improve accuracy, with diminishing returns to extra experts or methods. The gains are generally much greater from adding experts than from adding methods, and restricting the set of experts to those who do particularly well individually leads to the greatest improvements in the averaged assessments. The variability of assessment accuracy decreases considerably as the number of experts or methods increases, implying a large risk reduction. We discuss conditions under which the general pattern of results obtained here might be expected to be similar or different in other situations with multiple experts and/or multiple methods.

Key words: combining forecasts; correlation assessment; assessment methods; expert judgment

History: Received on May 22, 2002. Accepted by Don N. Kleinmuntz on October 31, 2002, after 2 revisions.

1. Introduction

Expert judgments can provide useful information for forecasting and decision making. In decision and risk analysis, experts are often the only available source of information about some important variables or relationships. When expert input is desired, the analyst must decide which expert(s) to consult and which method(s) to use to obtain information such as estimates or probabilities. The literature on combining forecasts (e.g., Clemen 1989) shows that obtaining multiple forecasts and combining them leads to improved forecasting performance, and that simple averages of forecasts seem to work well in comparison with more complex combining techniques. In decision analysis, a common procedure when assessing an expert's judgments is to use multiple assessment methods with the idea of having the expert look at the situation from a number of vantage points. This is usually done with the intention of ferreting out inconsistencies (e.g., von Winterfeldt and Edwards 1986). It can also provide multiple judgments from the same expert, and these judgments can be averaged just as we average judgments from multiple

experts. Makridakis and Winkler (1983) study averages of forecasts from different forecasting methods; forecasting accuracy improves and the variability of accuracy decreases as the number of methods increases. More generally, Brown and Lindley (1986) and Lindley (1986) point out that plural analysis, or plural evaluation (the use of multiple methods, models, or approaches), can be helpful at any stage in a decision analysis or for the entire analysis (e.g., structuring a problem in multiple ways).

The motivation behind using multiple experts and/or multiple methods is simply to get additional information that can lead to more accurate forecasts or estimates and, ultimately, to better decisions. The objective of this paper is to investigate the comparative benefits of using multiple experts, multiple methods, or some combination thereof. We use a data set from Clemen et al. (2000) involving correlation assessments by 90 subjects using 6 assessment methods. The experiment is summarized in §2, and averages of correlations from multiple experts (for the same method) and from multiple methods (for the same expert) are analyzed in §3. We consider the use of both multiple experts *and* multiple methods in §4. The

results are summarized and some concluding comments are presented in §5, along with some discussion of factors that could lead to similar or different results in other settings with multiple experts and/or multiple methods.

2. The Experiment

We summarize the important aspects of the experiment here. For complete details, see Clemen et al. (2000). The subjects were 90 Weekend Executive MBA students at the Fuqua School of Business at Duke University. All students had taken the statistics core course, which included probability, regression, and correlation, in the previous term. At the time of the experiment, they were taking the decision models core course, studying Monte Carlo simulation and the problem of modeling correlated variables.

Each subject assessed correlations for five pairs of variables using six methods. For each pair of variables, we provided the subjects with histograms of the marginal distributions and estimated means, standard deviations, and deciles of the distributions. No scatterplot or other indication of the joint distribution was provided, however, because we were interested in the subjects' ability to assess the level of dependence between the variables based on their own knowledge. The five pairs of variables and estimates of their Spearman correlations r (based on data obtained separately from the assessments) are:

(1) *Height and Weight (HW)*: The height and weight of male students enrolled in Fuqua's daytime MBA program (estimated $r = 0.530$ based on a sample of 218 students).

(2) *Math and Verbal SAT Scores (MV)*: Scores on the two portions of the Scholastic Aptitude Test (SAT) for Duke undergraduates (estimated $r = 0.377$ based on data for 46,278 Duke undergraduates who matriculated between 1963 and 1978).

(3) *Standard and Poor's 500 and Dow-Jones Industrial Average (SD)*: Monthly returns for these two indexes of stock market performance (estimated $r = 0.950$ based on monthly data from February 1945 to January 1995).

(4) *Eli Lilly and Chrysler Corporation (LC)*: Monthly returns for Eli Lilly (a pharmaceutical firm) and Chrysler Corporation common stock (estimated

$r = 0.173$ based on data from July 1976 through June 1996).

(5) *Automotive Index and Chrysler Corporation (AC)*: Monthly returns for an index of stocks (defined by Compustat) from the automotive industry and Chrysler Corporation common stock (estimated $r = 0.738$ based on data from July 1976 through June 1996).

For each correlation-assessment method, we provided some guidance as to how the subject might think about the question. The following six methods are illustrated with the assessment questions asked for *HW*:

(1) *S* (strength of relationship): "How would you characterize the strength and nature of the relationship between height and weight for male MBA students?" The response was given on a continuous line scale for which 1 = "very strong negative relationship," 4 = "no relationship," and 7 = "very strong positive relationship."

(2) *R* (correlation): "What would you estimate for the correlation between height and weight for male MBA students?"

(3) *CF* (conditional fractile): "A randomly chosen male MBA student is 74 inches tall, which is the 90th percentile of the height distribution. (This means that 90% of individuals are less than or equal to 74 inches tall.) For the same student, what percentile would you estimate for his weight?"

(4) *CNC* (concordance probability): "Suppose we randomly choose two male MBA students and label them *A* and *B*. Given that *A* is taller than *B*, what is your probability P_C that *A* also weighs more than *B*?"

(5) *JP* (joint probability): "A male MBA student has been randomly chosen. What is your probability that this student's height and weight are both in the lower 30% of their respective distributions (height less than or equal to 69 AND weight less than or equal to 160 pounds)? To put it more formally: What is $P(\text{height} \leq 69 \text{ and weight} \leq 160)$?"

(6) *CP* (conditional probability): "A randomly chosen male MBA student is in the lower 60% of the height distribution (height ≤ 71 inches). Given this, what is your probability that this student's weight falls in the lower 60% of the weight distribution (weight ≤ 175 pounds)?"

Each subject used all six methods for a given pair of variables before moving on to the next pair. The demographic pairs (*HW* and *MV*) were done first, followed by the financial variables (*SD*, *LC*, and *AC*), with randomized ordering within each of these groups. As for the methods, *S* was always asked first and the ordering of the remaining methods was varied from subject to subject using a Latin square design. For a given subject, the ordering of methods was the same for all variable pairs.

For analysis purposes, the assessments from the different methods were all converted to the following equivalent Spearman correlations:

- *S* was transformed linearly: $r_S = (S - 4)/3$.
- *R*: The response was taken to be r_R .
- *CF*: Using a nonparametric regression representation, we solved $E[F(X) | y] = r_{CF}[G(y) - 0.5] + 0.5$ for r_{CF} , where the question specified $G(y)$, and $E[F(X) | y]$ was the response.
- *CNC*: We assumed bivariate normality and found Kendall's $\tau = 2P_C - 1$, where P_C was the response. We converted τ to the Pearson correlation $r^* = \sin(\pi\tau/2)$ and solved $r^* = 2\sin(\pi r_{CNC}/6)$ for r_{CNC} .
- *JP* and *CP*: We assumed bivariate normality and calculated r_{JP} and r_{CP} accordingly.

The average and standard deviation of the equivalent correlations across subjects for each assessment method and variable pair are presented in Table 1 of Clemen et al. (2000, p. 1106), along with comparisons of these equivalent correlations with the corresponding data-based estimates.

For each equivalent correlation, the estimation error is the difference between that correlation and the data-based estimate. Measuring accuracy in terms of mean absolute error (MAE), Clemen et al. (2000) found that direct assessment (*R*) is the most accurate method, followed in order by *S*, *CNC*, *CP*, *CF*, and *JP*. In addition, directly assessed correlations show less variability than correlations derived from other assessment methods (followed by *S*, *CF*, *CP*, *CNC*, and *JP*) and yield less variable absolute errors (followed by *S*, *CF*, *CNC*, *JP*, and *CP*). *R*, *S*, and *CNC* had no mathematically inconsistent responses; *CF* had 5, *CP* 7, and *JP* 24 (out of 450). *S* was judged the easiest to use, followed by *R*, *CNC*, *CF*, *CP*, and *JP*. These results indicate that simply asking for a correlation is a reasonable approach, the simple strength-of-relationship method (*S*) performed quite well also,

and the method involving joint probabilities (*JP*) performed worst in terms of both accuracy and ease of use. Nonetheless, the accuracy of the better methods (including *R*) is somewhat limited. Using multiple experts or multiple methods is a promising strategy for improving accuracy.

3. Using Multiple Experts or Multiple Methods

Clemen et al. (2000) investigated a simple average of the equivalent Spearman correlations from all six methods and found this average to be more accurate than the best single assessment method. In this section, we consider such averages for two to five methods to study the impact of using multiple methods. For instance, for two methods, we calculated the average correlations for all $\binom{6}{2} = 15$ possible pairs of methods for each subject and each of the five variable pairs. For any given variable and expert, we ignored pairs of methods when one or both of the methods had missing correlations. Fortunately, only 65 (2.4%) of the 2,700 possible correlations (90 subjects \times 6 methods \times 5 variables) were missing and the missing data were spread across 28 subjects, so this was not a serious problem. We then investigated the accuracy of these average correlations for each pair of methods and each variable pair, using MAE as our primary measure of accuracy.

To investigate how averages from multiple experts perform, we went through the same procedure using averages across two to five subjects instead of averages across two to five methods. With two experts, this means calculating the average correlations for all $\binom{90}{2} = 4,005$ possible pairs of subjects for each method and each of the five variable pairs. With more experts, the number of possible combinations increases (to $\binom{90}{3} = 117,480$ for three subjects, for example).

Figure 1 shows the overall average MAE for averages of k experts and averages of k methods. As expected, the average MAE decreases as k increases for both experts and methods. The striking feature of Figure 1 is the much better performance from multiple experts than from multiple methods. From the starting point of one expert and one method, adding a second expert (using the same method) reduces MAE more than adding four more methods (with the same

Figure 1 Average MAE for Averages of 1–5 Experts and 1–5 Methods

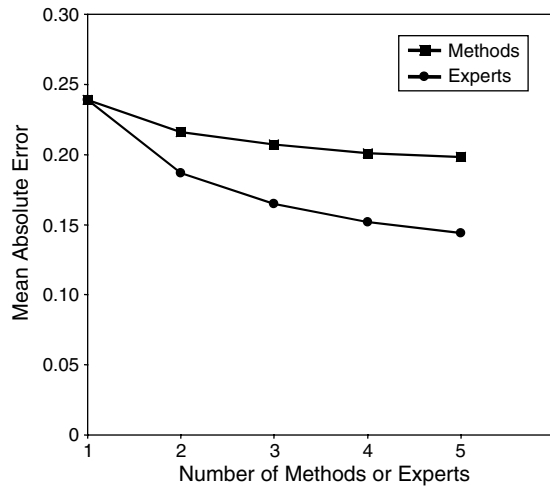
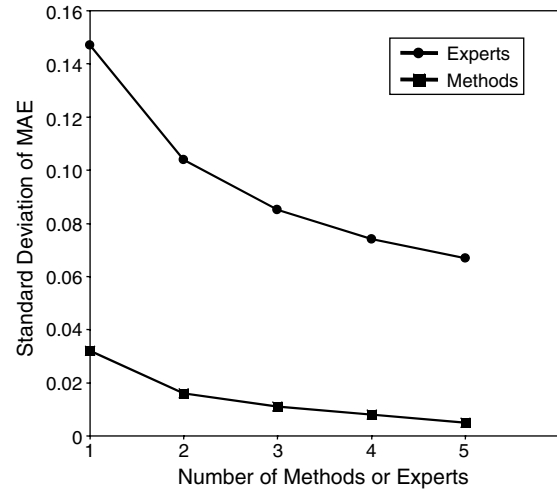


Figure 2 Standard Deviation of MAE for Averages of 1–5 Experts and 1–5 Methods



expert). The MAE for five experts is 75% of the MAE for five methods and 60% of the MAE for $k = 1$ expert or method. In Table 1, we break this down by variable pair and see that the dominance of multiple experts holds for all five pairs. For *MV*, *SD*, and *LC*, an average of two experts gives a lower MAE than an average of five methods; for *HW* and *AC*, two experts are not quite as good as five methods, but three experts are better than five methods.

The values in Figure 1 and Table 1 are averages across combinations of experts or methods, and the variability around these averages is also of interest. For each variable pair and $k = 1, \dots, 5$, we found the standard deviation of MAE across all combinations of k experts or k methods and then averaged these standard deviations across the five variable pairs. The resulting standard deviations are shown in Figure 2, where we see that the standard deviation of MAE is much higher for experts than for methods and decreases rapidly as k increases. The decrease is much greater in magnitude for multiple experts than for

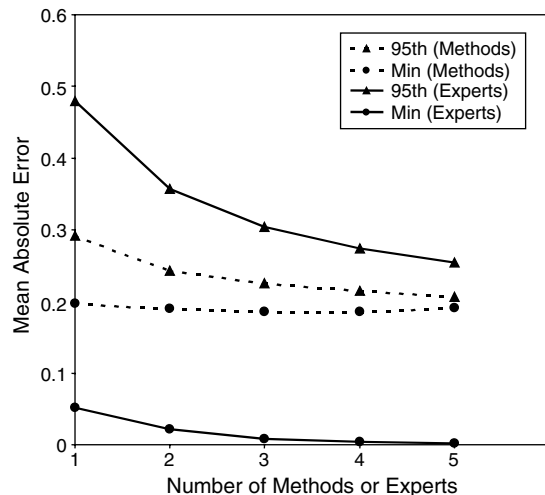
multiple methods, but greater in relative magnitude for methods (decreasing by a factor of roughly k^{-1}) than for experts (decreasing by a factor of roughly $k^{-1/2}$).

The higher standard deviation of MAE for experts (single or multiple) than for methods indicates an opportunity to get considerable improvements in performance with the better experts or combinations of experts. Of course, the other side of the coin is a greater risk of worse performance with the poorer experts or combinations. This is illustrated in Figure 3, which shows the minimum MAE and an approximate 95th percentile of MAE (the average MAE + 1.64 standard deviations) for all combinations of k experts or k methods, averaged across the five variable pairs. For both experts and methods, the upper values rapidly decrease as k increases. This decrease in the risk of poor performance is similar to results with averages of forecasts from forecasting methods in Makridakis and Winkler (1983) and is more dramatic for averages

Table 1 Average MAE by Variable Pair for Averages of 1–5 Experts and 1–5 Methods

Number of experts or methods	HW		MV		SD		LC		AC	
	Experts	Methods	Experts	Methods	Experts	Methods	Experts	Methods	Experts	Methods
1	0.096	0.096	0.125	0.125	0.096	0.096	0.091	0.091	0.119	0.119
2	0.055	0.069	0.065	0.096	0.062	0.080	0.048	0.067	0.072	0.089
3	0.034	0.060	0.046	0.087	0.051	0.076	0.034	0.060	0.056	0.078
4	0.026	0.055	0.036	0.083	0.046	0.074	0.028	0.056	0.048	0.072
5	0.022	0.052	0.030	0.081	0.042	0.073	0.024	0.072	0.043	0.068

Figure 3 Minimum and 95th Percentile of MAE for Averages of 1–5 Experts and 1–5 Methods



of experts than for averages of methods. For example, the 95th percentile decreases by 26% (47%) in going from one to two (five) experts and only 17% (29%) in going from one to two (five) methods. At the lower end, the best combinations of experts give incredibly low values of MAE (0.052 for $k = 1$, 0.022 for $k = 2$, and down to 0.002 for $k = 5$). Averages of methods provide less potential, with the lowest minimum MAE being 0.186. Of course, the greater number of experts (90) compared with methods (6) and the higher standard deviation of MAE for experts provide many more chances at very low MAE values as well as high MAE values for experts.

The motivation for using multiple experts or methods is obtaining additional information in order to get more accurate assessments. The greater improvement in MAE with multiple experts suggests that we get more “new information” from consulting an additional expert than we do from asking the same expert to think about the situation in a different way (i.e., to use a different method). The higher standard deviation of MAE for multiple experts is consistent with this in the sense that adding an expert is more likely to change the MAE by changing the average correlation. On average, the standard deviation of equivalent correlations is 0.23 across experts for a given method and 0.06 across methods for a given expert.

Another way of looking at the issue of information is to analyze the dependence among estimation

errors. For example, take the variable pair *HW*. Each expert used six methods to assess (directly or indirectly) the correlation for *HW*, yielding six assessed correlations and six corresponding estimation errors. With these six estimation errors for two experts, we calculated the correlation between the two experts’ errors for assessing *HW*. We found this error correlation for all 4,005 possible pairs of experts, summarized these 4,005 error correlations with their average (0.08) and standard deviation (0.46) as given in Table 2, and repeated the process for the other variable pairs.

Similarly, each method was used by 90 experts to assess the correlation of *HW*. For a specific method, we calculated the 90 estimation errors as above. With these 90 errors for two methods, we found the correlation between the two methods’ errors for assessing *HW*. We did this for all 15 possible pairs of methods, summarized the 15 correlations with their average (0.54) and standard deviation (0.08) in Table 2, and did the same for the other variable pairs.

Table 2 shows that the error correlations between methods are much higher on average (0.52, with a standard deviation of 0.12) than between experts (0.06, with a standard deviation of 0.48). In other words, the assessments using different methods for the same expert are somewhat redundant, and this redundancy tends to be consistent across experts. Assessments from different experts for the same method are nearly uncorrelated on average, but the error correlations are widely dispersed, implying that the strength and direction of the linear relationships among the subjects’ assessments vary greatly.

The average MAEs in Figure 1 and Table 1 are based on all possible combinations of experts or methods. How much could accuracy be improved by restricting attention to “better” experts or methods

Table 2 Averages (Standard Deviations) of Estimation Error Correlations

	Methods	Experts
HW	0.54 (0.08)	0.08 (0.46)
MV	0.56 (0.09)	0.11 (0.48)
SD	0.55 (0.13)	0.02 (0.49)
LC	0.50 (0.13)	0.04 (0.48)
AC	0.45 (0.17)	0.07 (0.48)
Averages	0.52 (0.12)	0.06 (0.48)

Table 3 Average MAE for Averages of 1–5 Methods and 1–5 Experts from Restricted Sets of Methods and Experts

	Number of methods or experts				
	1	2	3	4	5
R included	0.212	0.207	0.202	0.199	0.197
JP excluded	0.234	0.211	0.202	0.196	0.193
R in, JP out	0.212	0.206	0.199	0.195	0.193
JP included	0.269	0.225	0.212	0.204	0.200
All methods	0.240	0.216	0.207	0.201	0.198
Best 10% of experts	0.143	0.108	0.095	0.087	0.082
Worst 10% of experts	0.386	0.320	0.294	0.280	0.272
All experts	0.240	0.187	0.165	0.152	0.144

and/or eliminating “weaker” experts or methods? From §2, *R* is the best method, and *JP* is the worst method. Average MAEs are given in Table 3 for all possible combinations of methods and for combinations of methods restricted to include *R*, to exclude *JP*, to both include *R* and exclude *JP*, and to include *JP*. The restrictions to include *R* and/or exclude *JP* all reduce average MAE in comparison with the unrestricted MAE for the same number of methods, with the largest percentage reduction for two or more methods being about 5%. Increases in MAE due to considering only combinations including *JP* are of the same magnitude as decreases in MAE when considering only combinations including *R*.

With experts, we ordered the experts in terms of average MAE and then considered all possible combinations of the best 10% (9 of 90) and all possible combinations of the worst 10%. This resulted in dramatic changes in average MAE, as shown in Table 3. Restricting attention to the best 10% reduces average MAE by 40%–43% in comparison with the unrestricted MAE for the same number of experts. At the other end, the average MAE for the worst 10% is 61%–89% higher than the unrestricted MAE for the same number of experts. Somewhat surprisingly, for both the top and bottom groups, the percentage changes relative to the unrestricted group increase as the number of methods increases. This indicates the upside potential associated with being able to identify the better experts and average their assessments; the average MAE of 0.082 for combinations of five experts from the top 10% is 66% lower than the average MAE for $k = 1$ across all experts. Finally, although we might expect that the top experts

would be more similar to each other than to the other experts, the averages (standard deviations) of the error correlations between experts are roughly the same within the top group [0.13 (0.44)], within the bottom group [0.09 (0.45)], within the combined top and bottom groups [0.07 (0.45)], and for the entire set of 90 experts [0.06 (0.48)].

What about the question implied by the title of this section: Are we better off using multiple methods or multiple experts? On average, the gains in accuracy are much greater with multiple experts. The greater variability of MAE with multiple experts works together with the lower average MAE to give us a much better shot at an extremely accurate MAE with multiple experts. However, the higher variability with multiple experts suggests some risk of a high MAE. On balance, Figure 3 suggests that the advantage of averaging k experts over averaging k methods increases as k increases, because the MAEs for methods settle in a narrow range in the upper end of the range of MAEs for experts. The analyses involving “better” and “weaker” experts or methods are not exactly parallel because we have so many more experts than methods; the best or worst method constitutes 17% of the methods, as compared with the analysis of the top and bottom 10% of the experts. Table 3 suggests, however, that to the extent that we might be able to identify “better” experts or methods and “weaker” experts or methods, there is more potential for improved performance on the expert side.

Another way to address the question of methods versus experts is via a head-to-head competition. Table 4 shows the percentage of times that an average from a randomly chosen set of k experts ($k = 1, \dots, 4$) yields a lower MAE than an average from a randomly chosen set of m methods ($m = 1, \dots, 5$). Except for the

Table 4 Percentage of Times Average of Experts Has Lower MAE Than Average of Methods

Number of methods	Number of experts			
	1	2	3	4
1	60.5	74.6	81.2	84.9
2	52.7	68.8	76.5	81.0
3	48.7	65.7	74.1	77.3
4	47.0	63.9	72.4	77.5
5	46.1	62.7	71.5	76.7

Table 5 Percentage of Times Average of Experts Has Lower MAE Than Best Average of Methods

Methods	Number of experts			
	1	2	3	4
R	50.0	66.6	74.8	79.4
R, S	47.4	63.8	72.6	77.6
R, S, CNC	42.4	61.4	70.2	74.0
R, S, CNC, CP	42.8	61.2	70.4	75.8
R, S, CNC, CP, CF	42.2	60.4	69.6	75.0

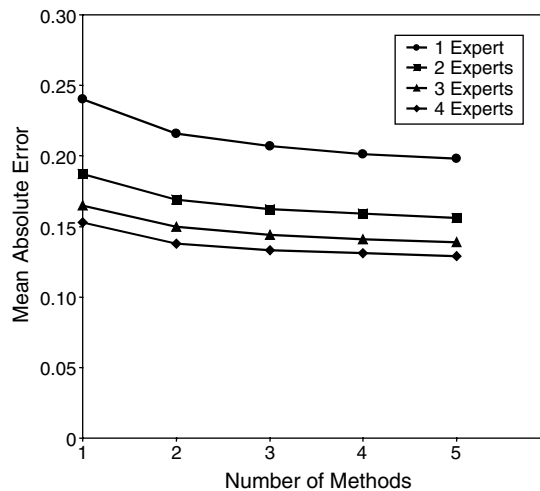
cases of one expert versus three, four, or five methods, where the percentages are slightly below 50%, the experts consistently win more often than the methods, sometimes by a large margin. For $k = m = 1, \dots, 4$, for example, the experts' winning percentages are 60.5%, 68.8%, 74.1%, and 77.5%, respectively. This verifies the above claim that as k increases, the advantage of averaging k experts over averaging k methods also increases. In all cases but one in Table 4, increasing both k and m by one increases the percentage of times the experts beat the methods. The exception is the slight decrease from 81.2% to 81.0% when going from $k = 3$ and $m = 1$ to $k = 4$ and $m = 2$.

The degree of superiority of averages of experts is shown in Table 5, which is identical to Table 4 except that for each value of m , the set of m methods with the lowest MAE is used. Interestingly, the optimal sequencing of methods is exactly in order of their individual MAEs. Even with the deck stacked in favor of the methods in this way, the experts still have the upper hand, and the percentages are not much lower than those in Table 4 for randomly chosen combinations of methods. For $k > 1$, a randomly chosen set of experts beats the best set of methods from 60.4% to 79.4% of the time.

4. Using Multiple Experts and Multiple Methods

If using multiple experts or multiple methods can be helpful, how about doing both? We found averages of equivalent correlations for all possible combinations of k experts and m methods, where $k = 1, \dots, 4$ and $m = 1, \dots, 5$. Figure 4 shows the overall average MAE for these combinations. Increasing k and/or m yields lower average MAE values, and it is clear from Figure 4 that there are diminishing returns to extra experts and to extra methods.

Figure 4 Average MAE for Averages of 1–4 Experts Using 1–5 Methods



At any point in Figure 4 with $k < 4$ and $m < 5$, we could consider adding another expert (moving down to the next lower curve) or adding another method (moving to the next point to the right on the same curve). In one case ($k = 3$ and $m = 1$), there is a slight advantage to adding another method. Note, by the way, that in this case adding another method means adding three data points because each of the three experts uses the new method; adding another expert means adding only one data point because the new expert will only have one method to use. In all other cases in Figure 4, we are always better off adding another expert, sometimes substantially better off. For example, when $k = 1$ and $m = 2$, adding another expert reduces MAE from 0.216 to 0.169, whereas adding another method gives an MAE of 0.207.

It is interesting to note that if we add both another expert and another method, the improvement in MAE is close to fully additive. For the 12 points in Figure 4 for which it is possible to add both another expert and another method, the improvement in MAE ranges from 92.2% to 100% of the sum of the separate improvements for adding just another expert and adding just another method, with an average of 96.2%.

5. Summary and Discussion

Our analyses of averages of correlations from multiple experts and/or multiple methods demonstrate a substantial degree of improvement in accuracy as we increase the number of experts or methods. For both

k experts and k methods, accuracy increases as k increases, with diminishing returns to extra experts or methods. The gains are much greater from multiple experts than from multiple methods, with a second expert being worth more on average than four more methods (starting from one expert using one method). Even at that, a second method reduces MAE by 10% on average and five methods provide a 17% reduction in MAE as compared with one method, so multiple methods can still be valuable. The comparable figures for two and five experts are 22% and 40%. In both cases, more than 50% of the improvement in going from $k = 1$ to $k = 5$ is obtained with the first addition. In simultaneously using multiple experts and methods, it is better to add another expert from most points with k methods and m experts, although adding a method can be preferable after adding a few experts.

The variability of MAE across different combinations decreases rapidly as k increases, implying a decrease in the risk of winding up with a “bad” combination. Indeed, this is a major benefit of combining. The variability of MAE is much greater for experts than methods, and pairwise error correlations are higher and less variable for methods. This suggests that there is more variability and less redundancy among the experts, which could explain why multiple experts perform better than multiple methods.

Our study focused on simple averages to combine the dependence estimates. More sophisticated combination procedures are available, but simple averages have been shown to perform well empirically and to be quite robust (e.g., Clemen 1989; Clemen and Winkler 1986, 1999). In addition, they are easy to use, requiring no assessments regarding the expert judgment process and no fitting. Based on experience with more complex combination procedures, we are not optimistic that they would improve performance more than adding another expert or method.

We analyzed an existing data set on dependence assessment. We acknowledge that the claims we can make on the basis of our analysis of one data set are limited, and we hope that our results will stimulate additional empirical studies to examine the implications of adding experts and/or methods in forecasting or other tasks. Such studies might focus, for example, on estimates, point forecasts, or probability assessments. Moreover, we hope our

results will stimulate further thinking and discussion of the underlying issues. In the remainder of this section, we initiate this discussion.

Expert Information and Accuracy. At a basic level, an expert has a set of information, including factors such as knowledge, experience, and specific data. A method is something the expert applies to that set of information to assess certain measures of interest. Different methods can lead to improvements in performance by causing an expert to consider different aspects of the information or to think about the same aspects in different ways, thereby teasing more out of the expert’s information. Nonetheless, no matter how many methods an expert uses, the results are limited by the expert’s personal set of information. In contrast, a new expert has a possibly distinctive set of information, which, at least in principle, could provide further improvement even if many experts have already been consulted. In this sense, using multiple experts has more potential for improving performance than does using multiple methods. This conjecture is supported by the minimum MAE curves in Figure 3 and the results for the top 10% of experts in Table 3.

When we start with a single expert using a single method, the focus is on the accuracy of different experts and methods. The term accuracy is used to refer to the size of the forecast errors, and MAE is the specific measure of accuracy used here. Other measures, such as mean square error, are frequently encountered. For probability forecasts, proper scoring rules (Winkler 1996) such as the quadratic or logarithmic rule are commonly used accuracy measures. The specific measure used should depend on the nature of the measures being combined and the underlying decision-making problem. Because we are not dealing with a specific decision-making problem, we use MAE as a generic penalty function.

Accuracy can be influenced by a number of factors. For instance, the error for a given expert using a specific method can be thought of as a sum of a bias term associated with the expert, a bias term associated with the method, and a random error term with mean zero. The benefits of averaging multiple forecasts arise from the usual averaging of random errors and also from any canceling of offsetting biases (which depends on the distribution of bias terms among

experts and methods). Note that in some instances, it may be desirable to use the same expert-method combination more than once to check on reliability, as in the maxim followed by carpenters and tailors to “measure twice, cut once.”

Dependence Among Experts or Methods. In going beyond one expert using one method, gains in accuracy will depend not only on the accuracy of the new expert or new method, but on the relationships among forecast errors. In terms of the error decomposition noted in the previous paragraph, dependence among errors can be caused by bias terms that do not cancel (representing common systematic biases) or by correlation among the random error terms (representing biases with respect to the specific assessment, such as might be caused by common information). It is well recognized in the literature on combining forecasts that a key determinant of the benefits of combining is the degree to which the errors from the individual forecasts are correlated. On the theoretical side, Winkler (1981) develops a model for combining information from dependent sources and Clemen and Winkler (1985) show how positive dependence can severely limit the gains from using additional sources. In a different context relating to decomposition in decision analysis, Ravinder et al. (1988) develop a model with dependent errors; Kleinmuntz (1990, p. 119) notes that “one of the most serious threats to decomposition’s error-reduction potential is the presence of dependent errors.” Empirical studies of forecast combination (Clemen 1989) show that forecast errors are often highly correlated. For recent related discussions of the role of dependence among information sources, see Soll (1999) and Ariely et al. (2000).

The low pairwise error correlations among the experts in our study are surprising. Perhaps more surprising is that these low correlations persist when we restrict attention to only the better experts (the top 10%) or the weaker experts (the bottom 10%). We believed the subjects would be reasonably knowledgeable about the pairs of variables considered in the experiment. However, they were not experts in the true sense of the word, and this may explain, in part, why their pairwise error correlations were so low on average and so variable. To the extent that those error correlations are higher and judgments less variable in a particular setting, the gains from multiple experts

would be reduced. We expect that in most settings with real experts, it will be hard to find low correlations because the experts are likely to have seen the same data, to have been exposed to the same theories and ideas, and so on. On the other hand, careful selection of experts with different viewpoints might temper this somewhat; in some cases (e.g., environmental risk assessment), experts have differed considerably in their judgments (e.g., Dewispelare et al. 1995, Morgan and Keith 1995). Expert selection is not a random process; the objective should be to choose the most useful set of experts. This generally means trying to select experts who are knowledgeable about the question at hand (implying accurate assessments) but differ from each other in the way they might think about that question to reduce redundancy.

Identifying Better Experts or Methods. In our analysis, Figure 3 suggests that if we could identify the best combinations of experts, we could get almost perfect accuracy, with MAE virtually zero. That may be infeasible, but a more reasonable task, narrowing the set of experts to those with better accuracy as we do in Table 3 with the top 10% of experts, yields substantial improvements in accuracy. Using just three experts from the top 10% reduces MAE by 60% as compared to using one expert from the full set. This result largely stems from the fact that even within the top 10%, pairwise error correlations are low on average. Table 3 further shows that adding fourth and fifth experts from among the top 10% improve MAE by 8% and 6%, respectively. In some circumstances, this could translate into a significant economic gain.

Because of the moderately high correlations among errors from the different methods and the low variability of MAE among the methods, the gains from identifying the better methods are not great. Nonetheless, it is still advantageous to use the better methods and avoid the weaker methods, and these results suggest that the development of new methods that have lower correlations with the current methods could be useful. However, low correlations may be difficult to achieve because an expert using two methods is constrained by a given set of information.

Sequentially Adding Experts or Methods. We can extend the notion of expert selection to the problem of sequentially adding components (experts or methods)

into the combination. The best component to enter the combination would typically have high accuracy and low dependence with previous experts and methods. Given experts with similar accuracy, which may often be the case in practice, adding the expert that is least correlated with the others makes sense. In our experience, however, we tend to find experts who have similar levels of dependence. The results of the current study suggest that, absent evidence to the contrary such as might be provided by some particularly low levels of dependence, experts should enter the combination in order of their individual accuracy. If we have a good set of experts from which to choose, this probably means that a number of experts will be added before a second method. When should we stop adding experts or methods? That depends on the situation. Clearly, there are diminishing returns, and most of the benefit occurs with the first three to four experts and/or methods, as suggested by previous research (Soll 1999).

Our large number of experts made a careful study of the optimal sequencing of experts impractical. We can, however, consider the optimal sequencing of methods. The ordering of our methods from best to worst, in terms of overall individual MAE, is *R*, *S*, *CNC*, *CP*, *CF*, and *JP*. From Table 5, that is just the way they enter the best combinations of methods as we increase the number of methods. The pairwise correlations among errors from the different methods are similar, ranging from 0.42 to 0.62, so there are no pairs of methods with particularly low dependence that could be used to reduce redundancy and improve the accuracy of the combination.

Of course, in practice, marginal costs as well as marginal benefits of adding experts or methods should be considered. In general, we would expect that extra experts would be more expensive than extra methods. Going from one to two methods could yield a relatively cheap but nontrivial improvement in accuracy. If increased accuracy is important, though, obtaining additional expert assessments may be worth the cost. As in any similar choice, the decision maker must assess the trade-offs between the extra costs and potential gains.

Acknowledgments

This research was supported in part by the National Science Foundation under Grant numbers SBR 98-18855 and SES 00-84383. The authors are grateful to Gregory W. Fischer for his collaboration on the original experiment, to Ilia Tsetlin for his help with the data analysis for the combined correlation assessments, and to the referees and Jack Soll for helpful suggestions.

References

- Ariely, D., W. T. Au, R. H. Bender, D. V. Budescu, C. B. Dietz, H. Gu, T. S. Wallsten, G. Zauberman. 2000. The effects of averaging subjective probability estimates between and within judges. *J. Experiment. Psych.: Appl.* **6** 130–147.
- Brown, R. V., D. V. Lindley. 1986. Plural analysis: Multiple approaches to quantitative research. *Theory Decision* **20** 133–154.
- Clemen, R. T. 1989. Combining forecasts: A review and annotated bibliography. *Internat. J. Forecasting* **5** 559–583.
- Clemen, R. T., R. L. Winkler. 1985. Limits for precision and value of information from dependent sources. *Oper. Res.* **33** 427–442.
- Clemen, R. T., R. L. Winkler. 1986. Combining economic forecasts. *J. Bus. Econom. Statist.* **4** 39–46.
- Clemen, R. T., R. L. Winkler. 1999. Combining probability distributions from experts in risk analysis. *Risk Anal.* **19** 187–203.
- Clemen, R. T., G. W. Fischer, R. L. Winkler. 2000. Assessing dependence: Some experimental results. *Management Sci.* **46** 1100–1115.
- Dewispelare, A., L. Herren, R. T. Clemen. 1995. The use of probability elicitation for high-level nuclear waste regulation. *Internat. J. Forecasting* **11** 5–24.
- Kleinmuntz, D. N. 1990. Decomposition and the control of error in decision-analytic models. R. M. Hogarth, ed. *Insights in Decision Making: A Tribute to Hillel J. Einhorn*. University of Chicago Press, Chicago, IL, 107–126.
- Lindley, D. V. 1986. The reconciliation of decision analyses. *Oper. Res.* **34** 289–295.
- Makridakis, S., R. L. Winkler. 1983. Averages of forecasts: Some empirical results. *Management Sci.* **29** 987–996.
- Morgan, M. G., D. W. Keith. 1995. Subjective judgments by climate experts. *Environ. Sci. Tech.* **29** 468A–476A.
- Ravinder, H. V., D. N. Kleinmuntz, J. S. Dyer. 1988. The reliability of subjective probabilities obtained through decomposition. *Management Sci.* **34** 186–199.
- Soll, J. B. 1999. Intuitive theories of information: Belief about the value of redundancy. *Cognitive Psych.* **38** 317–346.
- von Winterfeldt, D., W. Edwards. 1986. *Decision Analysis and Behavioral Research*. Cambridge University Press, Cambridge, U.K.
- Winkler, R. L. 1981. Combining probability distributions from dependent information sources. *Management Sci.* **27** 479–488.
- Winkler, R. L. 1996. Scoring rules and the evaluation of probabilities. *Test* **5** 1–60.